

DOCUMENT RESUME

ED 323 375

CE 055 770

AUTHOR Best, Fred; And Others
TITLE Quality Standards and Performance Measures: Policy Option Paper on Strategic Recommendation 7. Draft.
INSTITUTION Adult Education Inst. for Research and Planning, Sacramento, CA.; Pacific Management and Research Associates, Sacramento, CA.
SPONS AGENCY California Community Colleges, Sacramento. Office of the Chancellor.; California State Dept. of Education, Sacramento. Youth, Adult and Alternative Educational Services Div.
PUB DATE 17 Aug 90
NOTE 121p.; For related documents, see CE 055 768-779.
PUB TYPE Viewpoints (120)
EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Accountability; *Adult Education; Community Colleges; Performance Factors; *Performance Tests; *Program Evaluation; Quality Control; Skill Analysis; Standards; *Student Evaluation; Test Reliability
IDENTIFIERS 353 Project; *California; Strategic Planning

ABSTRACT

Addressing the implications of quality standards and performance measures for California's adult education system, this document focuses on skill assessment and testing systems as a core performance measure. Following an executive summary, the first chapter is an introduction that includes an overview of recommendation 7 of the 1989 Strategic Plan for Adult Education, which spurred the creation of this document. Chapter 2 provides background on student skill assessment and testing systems. Chapter 3 covers the criteria for evaluating skill assessment and testing systems. Chapter 4 contains 14 proposals that will lead to the development of quality standards and performance measures for adult education. Chapter 5 provides a summary of the future actions in research and planning needed to implement the proposals made in this document. The document concludes with 50 references. (CML)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED323375

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
- ☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

[Signature]

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Quality Standards and Performance Measures:

Policy Option Paper on Strategic Recommendation 7

Submitted To:

Adult Education Unit
Division of Youth, Adult and Alternative Educational Services
CALIFORNIA STATE DEPARTMENT OF EDUCATION
Sacramento, California 95814
(916) 322-2175

Chancellor's Office
CALIFORNIA COMMUNITY COLLEGES
Sacramento, California 95814
(916) 322-6880

Prepared By:

Fred Best, Ph.D.
Jon Ebeling, Ph.D.
Frederica Shockley, Ph.D.
Van Woolley, M.A.

ADULT EDUCATION INSTITUTE FOR RESEARCH AND PLANNING
429 J Street, Suite 400
Sacramento, California 95814
(916) 448-9063

August 17, 1990

Letter of Transmittal

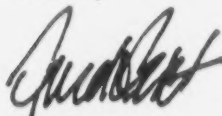
This draft report has been prepared as the first of three planning papers in response to Recommendation 7, *Quality Standards and Performance Measures*, of California's 1988-89 *Strategic Plan for Adult Education*. It has been prepared as a collaborative effort by staff at the Adult Education Institute for Research and Planning (AEI) and Jon Ebeling and Frederica Shockley of California State University of Chico.

During this period between May and July 1990, this report has been will be reviewed and revised by the Subcommittee on Quality Standards and Performance Measures of the 1990-91 Interim Adult Education Steering Committee. It is now under final review prior to formal submission to the Interim Steering Committee.

As in the past, future review will be conducted through public meetings, which will be open to observation to all concerned parties. Preliminary recommendations from this committee will be submitted to State Superintendent of Instruction Bill Honig, State Community College Chancellor David Mertes, and other participating agency heads. Pending acceptance or adjustment by participating agency heads, the recommendations of this report will be used to guide future research and planning to develop *Quality Standards and Performance Measures for Adult Education* within California.

Persons concerned with this issue are encouraged to contact the Adult Education Institute or Steering Committee members to share views and obtain additional information.

Sincerely,



Fred Best
Executive Director
Adult Education Institute for Research and Planning

STEERING COMMITTEE AND PROJECT STAFF

Interim Steering Committee

PAUL BELOMY
Principal
Santa Clara Adult School
Santa Clara

JAMES FIGUEROA
Assistant Superintendent
Adult/Occupational Education
Los Angeles USD
Los Angeles

NOREEN HANNA
Chair, ACSA
Adult Education Committee
Napa

PERCY JULIEN
Director, Adult Education
Oakland Unified School
District, Oakland

ANITA MARTINEZ
Representative, Academic
Senate for Calif Community
Colleges, San Francisco
Community College

BONNIE OLIVER
Director, Special Projects
KCET-TV Educational
Enterprise Division
Los Angeles

DICK RANES
Unit Manager
IBM
Costa Mesa

PRANY SANANIKONE
Director for Outreach,
UCI Extension,
University of California,
Irvine

GARI BROWNING
CATESOL Representative
Costa Mesa

JEROME FINNIGAN
Education and Training
Xerox Corporation
El Segundo

GREGORY HEARN
Superintendent
San Benito High School District
Hollister

DARLENE LAVAL
Board of Education
Fresno Unified School District
Fresno

SUZANNE McCORMICK
Vice President, California
Manufacturers Association
Service Corporation, Sacramento

HENRY PAGE
Principal
Palo Alto Adult School
Palo Alto

JOE RICHEY
Executive Vice President
Industry Education Council
of California, Woodland

ADRIANA SANCHEZ-ALDANA
ESL/Amnesty Resource Teacher
Sweetwater Union
High School District, Del Rey
Adult Center, Chula Vista

GARY E. STRONG
State Librarian
California State Library
Sacramento

CHARLES CARLSON
Dean of Instruction
Bakersfield College
Bakersfield

RONALD GRISTOCK
Vice President
Community Affairs
Wells Fargo Bank
San Francisco

AUGUST HUESCA
Provost, Adult Continuing Ed
North Orange County Community
College District, Fullerton

JERRY LEW
Regional Training Coordinator
Kaiser Permanente
Regional HRD
Oakland

DANIEL McQUILLEN
Director
Kern County Regional Occupational
Program, Office of Kern County
Superintendent of School, Mojave

CORINNE PRICE
Director
Salinas Adult School
Salinas

LEONARD RIVERA
Principal
Whittier Adult School
Whittier

ROD SMITH
Provost, Educational Cultural
Complex, San Diego Community
College District, San Diego

WILLIAM WENRICH
Chancellor
San Diego Community College
District, San Diego

LANI DEVINCENTIS
Dean of Noncredit Education
Glendale Community College
Glendale

SANDRA HAMILTON
Executive Director
Southeast Los Angeles County
Private Industry Council, Inc.
Cerritos

THOMAS JOHNSON
Assistant Superintendent
Adult & Alternative Education
Hacienda La Puente USD
La Puente

HAZEL MAHONE
Superintendent
Grant Joint Union High School
District, Sacramento

CONSUELO NIETO
Professor
Graduate School of Education
CSU, Long Beach

WALTER POPKIN
President, California Council
for Adult Education
Montebello

FRANCES RUIZ
ESL Teacher
Corona-Norco Adult School
Corona

JUANITA STANLEY
Executive Director
California Literacy, Inc.
San Gabriel

State Education Staff

DOUG BURRIS
Deputy Chancellor
Chancellor's Office
California Community Colleges

RITA CEPEDA
Dean, Education Standards &
Evaluation, Chancellor's Office
California Community Colleges

GERALD KILBERT
State Director
Adult Education
California Dept of Education

RAY EBERHARD
Program Manager
Adult Education Unit
California Dept of Education

NORMA MORRIS
Analyst, Noncredit & Community
Services Chancellor's Office
California Community Colleges

KATHY WARRINER
Analyst, Basic Skills & Literacy
Chancellor's Office
California Community Colleges

LYNDA SMITH
Project Monitor
Adult Education Unit
California Dept of Education

Institute Staff

FRED BEST, Executive Director
CUBA MILLER, Senior Associate
BERNADETTE DAWSON, Associate
STEVE PORTLOCK, Business Manager
ANAN NUGGIBALLI, Computer Systems



JIM McDONALD, Publications and
Public Information
TAMMY LEWIS, Research Assistant
DENNIS PORTER, Consulting Specialist
VAN WOOLLEY, Senior Consultant

Adult Education Institute for Research and Planning
Sacramento, California

Contents

<i>Letter of Transmittal</i>	<i>i</i>
<i>Advisory Committee and Project Staff</i>	<i>ii</i>
<i>Acknowledgements</i>	<i>iv</i>
<i>Executive Summary</i>	<i>v</i>
1. Introduction and Overview	1
2.1 History and Summary of Recommendation	2
2.2 Focusing the Recommendation	5
2.3 Linking Standards and Measures	8
2.4 Steps to Implement Recommendation	9
2. Background on Skill Assessment Tests	11
2.1 Testing in the Context of Changing Society	11
2.2 What Do Stakeholders Want from Testing Systems?	14
2.3 An Overview of Skill Testing Systems	17
2.4 Considerations and Developments	23
2.5 Forging Skill Assessment Tests for Tomorrow	31
3. Standards for Skill Assessment Tests	33
3.1 Availability of Technical Documentation	34
3.2 Applicability of Test to Multiple Uses	36
3.3 Pertinence of Scores and Range of Measurement	37
3.4 Documentation of Validity and Reliability	44
3.5 Representative and Comparative Data	48
3.6 Common Measurement of Skill Levels	60
3.7 Capacity for Analysis and Reporting	62
3.8 Protection of Individual Rights and Privacy	65
3.9 Cost Effectiveness and Administrative Viability	66
3.10 Technical Auditing and Test Development	68
4. Proposals	70
4.1 Priorities and Uses	70
4.2 Criteria for Skill Testing Systems	74
5. Future Action	84
5.1 General Review	84
5.2 Developmental Considerations	85
5.3 Future Actions	86
Bibliography	87

Acknowledgements

This paper was prepared with advice and information from many people. The Adult Education Institute for Research and Planning (AEI) appreciates the assistance provided so generously from the following persons.

- | | |
|--|--|
| Mr. Tom Bauer, State Department of
Education, Sacramento, California | Dr. Thomas G. Sticht, Applied Behavioral
and Cognitive Sciences, San Diego,
California |
| Ms. Norma Carolan, State Department of
Education, Sacramento, California | Ms. Cuba Miller, Adult Education Institute,
Sacramento, California |
| Ms. Bernadette Dawson, Adult Education
Institute for Research and Planning,
Sacramento, California | Dr. Sarah Newcomb, U.S. Department of
Education, Washington, D.C. |
| Dr. Ray Eberhard, State Department of
Education Sacramento, California | Dr. Stephen Reder, Northwest Regional
Labs, Portland, Oregon |
| Dr. Julia Goodwin, Program Development
Supervisor, State Department of
Education, New York | Mr. Harry Van Houten, Office of Program
Development, Division of Adult
Education, Stater Department of
Education New Jersey |
| Dr. Greg Jackson, Association of
Community-Based Education,
Washington, D.C. | Technical Sergeant Gregory Ward, 3564th
Air Force, Recruiting Squadron, "E"
Flight, Chico, California |
| Jim McDonald, Adult Education Institute
for Research and Planning, Sacramento,
California | Dr. Chris Zachariadis, Association of
Community Based Education,
Washington, D.C. |
| Staff Sergeant John Passanelli, U.S. Army,
Chico Recruiting Company,
Sacramento Recruiting Battalion | Ms. Kathy Warriner, Chancellor's Office,
California Community Colleges,
Sacramento, California |
| Dr. Ron Pugsley, U.S. Department of
Education, Washington, D.C. | |
| Dr. Bob Simmons, Director of Adult
Education Division, State Department
of Education, New Jersey | |

The contents of this paper do not necessarily reflect the views of the above persons or institutions they represent.

Executive Summary

This options paper outlines a criteria for selecting or developing quality standards and performance measures that ensure that learners, providers and taxpayers benefit from higher quality knowledge about what transpires in the adult education.

This Executive Summary encapsulates the first of three papers dealing with the implementation of quality standards and performance measures for California's adult education system. These papers are being prepared in response to a recommendation of the 1988-89 *Strategic Plan for Adult Education*.¹

This first paper focuses on *skill assessment and testing systems* as a core performance measure. The term *systems* is more appropriate to this discussion than *tests*. It communicates a more holistic view that encompasses how tests are administered, how they are interpreted and how these interpretations are used.

This first paper proposes generic criteria for developing these testing systems to meet a variety of needs. The second paper will be a feasibility study which will use the generic criteria to evaluate testing systems and will make recommendations regarding the appropriateness of existing assessment procedures, instruments and systems. The third paper will recommend a plan to implement quality standards and performance measures.

This first paper has five chapters dealing with: (1) Introduction and Overview of Recommendation, (2) Background on Student Skill Assessment and Testing Systems (3) Criteria for Evaluating Skill Assessment and Testing Systems, (4) Recommendations, and (5) Future Actions.

Introduction and Overview

Recommendation 7 of the 1988-1989 *Strategic Plan for Adult Education* calls for Program Quality Standards and Performance Measures. The exact text of this recommendation is provided below:

¹ *Adult Education for the 21st Century: Strategic Plan to Meet California's Long-Term Adult Education Needs* (1989 Edition), pages 64-65.

Recommendation 7
QUALITY STANDARDS
AND PERFORMANCE MEASURES

Develop, with local participation, Program Quality Standards and Performance Measures that are appropriate to adult learning.

Program Quality Standards should:

- Applicability to varied methods of adult education (e.g. tutoring, class instruction, TV courses, etc.).
- Provide minimum standards for certifying programs for funding and legal status as educational providers.
- Establish guidelines for excellence and criteria for funding incentives (e.g. quality standards, quality indicators and quality frameworks).

Performance Measures should:

- Provide standard assessment of the educational performance of individuals and providers.
- Emphasize outcomes and skill improvement rather than inputs and processes.
- Measure performance for a variety of educational endeavors and approaches.

These measures should not necessarily replace existing standards and measures, but provide a common standard that can be used by all providers.

Research and planning on Recommendation 7 should be prioritized before effective action can be taken.

DEFINING STANDARDS AND MEASURES

First, it is necessary to specify and elaborate the meaning of quality standards and performance measures:

- **Program quality standards** specify the expectations of stakeholders for programs. These standards can define minimum requirements, normal expectations and conditions of excellence.
- **Performance measures** provide a way of documenting whether programs meet or surpass quality standards. They can focus on student outcomes (e.g. achievement test scores, program completions) or program inputs (e.g. student-teacher ratios, teacher certification, dollars spent per student).

Both quality standards and performance measures can have many manifestations.

FOCUSING ON LEARNER ACHIEVEMENT

A key theme of Recommendation 7 is that quality standards and performance measures should focus on student outcomes—most notably, demonstrated improvement and mastery of specified knowledge and skills.

Without disavowing other standards and measures, learner attainment of knowledge and skills is one of the common denominators that can be prescribed, measured and compared across all programs.

This focus stemmed from a strong belief among the 1988–89 Adult Education Advisory Committee that the goals of instruction must be clearly specified and that outcomes be viewed in terms of what students actually learn. This theme does not suggest that program inputs (e.g. modes of instructions, facilities, teacher certification, dollars spent per student, curriculum content, etc.) and other measures of program performance are unimportant. Without disavowing other standards and measures, learner attainment of knowledge and skills is one of the common denominators that can be prescribed, measured and compared across all programs. While there are many ways to measure "student outcomes," deliberations of the 1988–89 Adult Education Advisory Committee focused on the instruments and procedures that compose learner assessment and skill testing systems.

PRIORITIZING BASIC SKILLS AND LITERACY

It is impossible to develop quality standards and performance measures for all these areas of adult instruction simultaneously. Priorities need to be specified to provide a manageable plan for developing them.² Recommendation 7 does not prioritize areas of instruction. However, student enrollment patterns coupled with social and economic needs clearly point to *adult literacy*, as the first and highest priority. Adult literacy includes basic skills such as reading, writing, verbal communication and numeracy.

This report will propose general criteria to guide the selection and development of skill testing and assessment systems. These criteria will be used initially in the areas of adult literacy and basic skills, but they can be applied to skill testing and assessment systems for other areas of instruction.

LINKING STANDARDS AND MEASURES

Another intent of Recommendation 7 is that performance measures be linked to quality standards. Specifically, quality standards should be established, then used to guide the development of performance measures. In short, we should have a way to measure whether we achieve what we intend to accomplish.

This will require coordination of Adult Education Institute (AEI) implementation research with a number of agency projects. Most notably, both the State Department of Education and the Community Colleges have ongoing projects to establish curriculum quality standards for programs:

² Many of these areas of instruction lend themselves to common standards and performance measures. However, several areas will require unique ones. Areas such as substantially handicapped, parent education, and older adults may not lend themselves to "paper and pencil tests" and assessment. They may require different types of assessment and testing.

STEPS TO IMPLEMENT RECOMMENDATION

Over the next two years, the Adult Education Institute for Research and Planning (AEI) will work with the Interim Adult Education Steering Committee and other projects to accomplish four steps to develop quality standards and performance measures:

Recommendation 7 does not prioritize areas of instruction. However, student enrollment patterns coupled with social and economic needs clearly point to adult literacy, as the first and highest priority.

- (1) Establish General Criteria for Selecting or Developing Skill and Assessment and Testing Systems.
- (2) Develop Program Quality Standards for Areas of Adult Instruction, with Basic Skills and Literacy Receiving Top Priority.
- (3) Use Quality Standards and the Criteria to Select or Develop Skill Assessment and Testing Systems.
- (4) Use Quality Standards to Select/Develop Authentic or Applied Performance Measures, as well as Paper-Pencil Tests.

This paper will deal with the first of the previous four steps.

Background on Skill Assessment and Testing System

Key background information concerning assessment and test systems includes the following five topics:

- (1) Testing and Changing Social Needs
- (2) Testing Needs of Stakeholders
- (3) Overview of Skill Assessment and Testing Systems
- (4) Considerations and New Developments
- (5) Importance of Testing Standards

These topics provide background for the planning and development of skill assessment and test systems that will meet minimal testing standards while serving a variety of needs.

TESTING IN THE CONTEXT OF A CHANGING SOCIETY

There is nothing new about testing. In its most rudimentary form, testing began with the early foundations of civilization, when human communities first achieved a body of knowledge and skills that had to be passed on from generation to generation, or person to person. Those responsible for imparting knowledge found ways to assess whether learners had acquired the necessary skills. This took the form of tribal rites of passage to test endurance or hunting skills. It evolved into the day-to-day watchfulness of the master over an apprentice and the verbal drills of a medieval tutor and his pupil.

Testing, as we commonly know it today, began with the introduction of orchestrated group learning such as classroom instruction. This mode of education created the need for a fair and consistent way of assessing the skills and knowledge of individuals learning in such groups.

The next step in the evolution of testing was the emergence of large scale standardized testing systems to determine the knowledge or skills of individuals across geographic regions and school districts. Testing systems evolved to achieve economies of scale, to permit comparison of individual results to a larger population group, establish measurable standards for certifying individuals for further education or employment, and to assess the performance of large groups or educational institutions. Emergence of these systems was also fueled by advances in data processing technology and the science of tests and measurements.

Standardized testing systems have not occurred without criticism and reservations. In a review of the strengths and weaknesses of standardized testing systems, the American Psychological Association concluded that all such tests are not always soundly developed or appropriate to the ways they are used, but that *"the proper use of well constructed and validated tests provides a better basis for making some important decisions about individuals and programs that would otherwise be available."*³

The basic purposes of testing has remained constant throughout history. One purpose of testing is to determine whether individuals or groups have achieved a desired level of skills or knowledge.

In broad overview, the basic purpose of testing has remained constant throughout history. One purpose of testing is to determine whether individuals or groups have achieved a desired level of skills or knowledge. We use tests to provide learners with feedback on how well they are doing and what they need to learn. We use tests as rite of passage to certify whether one is qualified to assume social responsibilities or pursue further education. We use tests increasingly to provide a fair and consistent way of determining whether individuals have achieved required skills.

These roles for testing will continue. However, it is likely that the uses of tests and testing systems may change and grow in response to new social and economic needs:

- Testing to Monitor the Nature and Level of Skills within Our Population
- Testing to Provide Consistent, Fair and Flexible Certification of Individual of Skills and Knowledge
- Testing to Plan and Evaluate Education and Training Programs by Measuring the Skill Attainment of Students
- Testing as a Learning Technique that Uses Feedback to Guide Students

The roles of testing and testing systems in the future are likely to become increasingly important and comprehensive. However, these roles are likely to be viewed differently by groups impacted by testing.

³ *Standards for Educational and Psychological Testing*, American Psychological Association, Washington, D.C., May 1986, page i.

WHAT DO STAKEHOLDERS WANT FROM TESTING SYSTEMS?

The stakeholders of tomorrow's skill assessment and testing systems will be learners, educators and taxpayers. Each group has specific needs. At the same time, there are many common themes that suggest the need to design a testing system that serves all needs in a cost-effective fashion (See Exhibit 1).

Exhibit 1 SKILL TESTING NEEDS OF STAKEHOLDERS

Needs of Adult Learners

- Assessment of Educational Needs for Placement
- A Learning Plan Based on Student Goals and Specific Skills Needed to Achieve That Goal
- Feedback on Learning Achievement
- Certification of Skill and Knowledge Attainment for Education and Employment

Needs of Educators:

- Planning Instruction with Feedback to Learner
- Educational Output Data to Evaluate and Guide Programs
- Fair Testing System
- Efficient and Cost-Effective Testing System

Needs of Taxpayers:

- Monitor Skill and Knowledge Attainment of Population
 - Program Performance Indicators and Cost-Effectiveness
 - Certification of Skills
-

AN OVERVIEW OF SKILL TESTING SYSTEMS

There are a multitude of skill assessment and testing systems. Many of these are not appropriate for adults. This section provides a context for examining assessment and testing systems.

Skill Assessment and Testing Systems within Adult Educational Settings

Skill assessment testing within adult educational settings is commonly administered through local jurisdictions to comply with state and federal requirements. For the most part, these testing systems are applied with limited control by state or federal agencies. This section provides an overview of such arrangements within California and other states.

Skill assessment and testing systems within California evolved for many reasons. Key factors include the need for economies of scale in the development and use of tests, federal and state law mandating comparable test measurements, the need to link assessment of learning to movements for "competency-based education", and growing pressures to measure educational attainment and the performance of instructional programs

Starting in the 1960's, statutory requirements began to push adult educators toward more standardized approaches to testing.

Prior to 1960, assessment and testing in the adult education system was restricted to the use of teacher-designed tests and existing testing systems. For the most part, there were few common testing requirements, and even fewer common testing measurements.

Starting in the 1960's, statutory requirements began to push adult educators toward more standardized approaches to testing. The result was an increase in testing and reported test scores, but little improvement in testing approaches or standardization of testing results.

Movement toward more systematic testing was encouraged by the "competency based adult education" (CBAE) movement. This approach to instruction stressed a *"performance-based process leading to demonstrated mastery of basic and life skills necessary for the individual to function proficiently in society."*⁴ As curriculum was developed for this approach it became apparent that the *"demonstrated mastery of basic and life skills"* aspect of the CBAE approach needed consistent measures of performance. This focused attention on the need for a basic skills assessment and testing system for adults.

In 1982, California adopted policies that required the use of a competency based approach in adult education. These policies required programs using federal *Adult Education Act* funds to implement a competency based approach within three years to continue to receive these funds. These programs included adult basic education (ABE), English-as-a-Second Language (ESL), and Vocational ESL (VESL).

Skill assessment was a key component of this 1982 mandate. Agencies that received Section 306 (now Section 321) funds were required to administer the Comprehensive Adult School Students Assessment System (CASAS) to a representative sample of their students.⁵ Additionally, a system of pre/post testing was applied to determine how well students learned skills between the beginning and end of courses. These pre/post tests were used to monitor student progress in courses dealing with Adult Basic Education (ABE), English-as-a-Second Language (ESL), and Vocational English-as-a-Second Language (VESL).

Use of CASAS has grown during the 1980's. CASAS has clearly become the leading skill assessment and testing system within California. However, it is important to recognize that it is not used universally. Many school districts and community colleges use other assessment tests for purposes of counseling, placement and certification. These include:

- MDTP (Math Development Test Program)
- TABE (Test of Adult Basic Education)
- ABEL (Adult Basic Education Levels)
- ACT (American College Test)
- Locally Developed Tests
- College Board Developed Tests

⁴ *Evaluation Study Report: Competency Based Adult Education in California*, CBAE Staff Development Project, San Francisco State University, San Francisco and CASAS Project, San Diego Community College District, San Diego, March 1987. See also, National Center for Educational Statistics, 1982, page 80, as quoted in Judy Alamprese, *Investing in Change: Competency-Based Adult Education in California*, California State Department of Education, March 1987, ERIC ED 289 010, page 8.

⁵ Judy Alamprese, *Op. Cit.*, page 13

During the late 1980's, pressures grew to use skill test data as a means of measuring program performance. This created a need to generate test data that is representative of all students participating in programs.

Increasing demands for accountability and better assessment make it necessary to carefully review the status of CASAS and other systems prior to moving forward in the development of quality standards and performance measures.

Other states have the same needs and problems with concerning skill assessment and testing systems. New York and New Jersey illustrate developments beyond California's boundaries. New York state recently began to collect demographic information and to test all adult learners with Test of Adult Basic Education (TABE). New Jersey has an adult basic skills testing system which is similar to New York's. Incoming students are tested within 10 hours using either the TABE or Adult Basic Education Levels (ABLE) reading comprehension tests.

Many barriers to the development of comprehensive skill assessment and testing systems stem from the inability to enforce common test measures and standard testing procedures.

The challenges of skill assessment testing within adult education settings are the result of multiple factors. First, American school systems have a traditional commitment to local control. This local autonomy makes it difficult to require use of a common test system and standardized testing procedures. Second, resources are scarce within adult education, and providers have several constraints concerning how much time and money they can commit to testing.

Centralized Testing Systems

Many barriers to the development of comprehensive skill assessment and testing systems stem from the inability to enforce common test measures and standard testing procedures. This section explores the conditions necessary to establish common tests and testing procedures by examining the highly-regulated skill assessment system of the military.

The Department of Defense (DOD), with its unique opportunity to regulate the behavior of its "students", operates a highly-regulated testing system. Since civilian teachers do not have as much influence on their students as drill sergeants exercise over their recruits, this system would not be as successful in adult educational settings. However, the model does provide a good example of rigorous and centralized testing standards for the guidance, training and evaluation of adult populations. Some of these practices might be used to improve civilian testing and teaching systems.

The DOD uses the Armed Services Vocational Aptitude Battery (ASVAB)—a highly developed predictive measure of intellectual aptitudes for military occupations. This type of testing, which relies on the works of Galton and Binet, has been used to screen incoming recruits since World War I.⁶

The DOD tests, trains and retests many times with the same instrument. Records have been maintained on individual recruits for most of this century to predict performance in military occupations. As a result, the military has excellent comparative data.⁷ Testing of recruits and potential recruits has resulted in one of the largest sets of data on intellectual aptitude

⁶ Thomas G. Sticht, "Cognitive Science and Human Resources Management," *Advances in Reading/Language Research*, (Greenwich, Connecticut, JAI Press, 1982), page 5.

⁷ This reference is to the ASVAB14 (ASVAB A Brief Guide for Counselors and Educators, U.S. Government Printing Office, Washington, D.C., 1986).

and performance found anywhere in the United States. The military can examine the impact of program changes on new recruits over time. At a given point in time they can compare the impact of various programs on recruits with similar background characteristics. Or the military can compare the impact of a program on groups of recruits with different background characteristics. Such systems of testing and analysis can provide high quality information for planning, programming, teaching and guidance.

While large, centralized systems such as the ASVAB may not be transportable to adult civilian populations. Some of the approaches used by such systems are worthy of further study. Advances in information technology are likely to enable civilian testing systems to provide many of the same functions as those of the military without a highly-regulated system. Specifically, the use of computer-assisted interactive instruction, decentralized data storage through devices such as the proposed *EduCard™*, and efficiently designed sampling techniques hold the promise of providing more consistent and comprehensive testing.

CONSIDERATIONS AND DEVELOPMENTS

There are a number of considerations and developments to consider in the planning of quality standards and performance measures. These include new or changing theoretical perspectives, new technologies, and changing social and economic conditions.

Theoretical Perspectives

Two important theoretical perspectives to consider include andragogy and views of literacy.

Testing and instruction should reflect the needs of adults. The instruction of adults has been named "andragogy" (the art and science of teaching adults). In pedagogy (teaching children), the learner plays a dependent role. The teacher determines the subject matter, the teaching methods and the testing procedures. In andragogy, the students who are self-directed have some control over the subject matter, the teaching methods, and the testing procedures. Andragogical principles have two implications for testing. First, testing should be self-paced and emphasize self-diagnosis. Second, test measurements and scales should be relevant to adult life.

Given that our highest priority for skill assessment tests deals with basic skills and literacy, it is important to note the existence of different schools of thought about literacy. Definitions of "literacy" vary notably. In terms of testing, there are three schools of thought:

- (1) Academic Definition
- (2) Functional Life-Skills Definition
- (3) Functional Job-Skills Definition

Before deciding whether to measure functional literacy, job literacy or general literacy, adult educators will want to consider the learner's goals, the rate at which literacy can be achieved, the value of the skills acquired, and the expected public support.

New Knowledge and Technology

New knowledge and technology are likely to change our expectations and options for adult assessment testing. Key issues to be considered include computer aided testing, new delivery systems, new measurement technology, and item-response theory.

Changing Social and Economic Needs

Ongoing social and economic changes will also impact the need for adult skill assessment and testing systems. Two key trends to consider include the skill gap and demographic changes.

In coming years, the roles of testing in our society are likely to grow and diversify. There will be a need to better monitor the skill level of our population, document the learning output of our educational systems, and provide better and fairer assessment and certification of individual skills.

A growing gap between the skill requirements of today's world and the skills of our population indicate that we must measure more effectively the skills of our society and provide feedback to individuals as to whether they have the skills needed to lead productive and rewarding lives.

The increasing diversity of California's population indicates that testing systems of all types must be applicable to persons from different backgrounds. For example, tests must be culturally neutral to measure the competencies of Hispanics as well as Anglos. Using computer based testing in some classes may alter the performance of some examinees by creating excessive test anxiety. Some students from societies without such technology may find the computer too much to manage while being tested too. ESL classes may be particularly susceptible to this problem. As a result, ESL students may require different literacy programs and different testing systems according to Sticht. Tests must also be as easily understood by senior citizens who may have difficulty seeing small print or hearing at normal voice levels.

FORGING SKILL ASSESSMENT TESTS FOR TOMORROW

In coming years, the roles of testing in our society are likely to grow and diversify. There will be a need to better monitor the skill level of our population, document the learning output of our educational systems, and provide better and fairer assessment and certification of individual skills. These needs will be matched with new technologies and techniques that will enable us to expand testing service by better integrating them into the fabric of our educational systems. Testing is a powerful tool for assisting individuals and society understand where we are and what we must do to improve learning, and ultimately to reinforce the actual learning process. These goals can be achieved. However, efforts must be based on a common agreement about our testing needs, understanding of minimal standards for testing, and commitment to marshal technology and resources to make testing a tool to enhance hope and opportunity within a context of accountability for learning.

Standards for Skill Assessment and Testing Systems

This chapter discusses ten standards, or criteria, for evaluating skill assessment and testing systems. These criteria are intended to serve as general standards for the selection or development of skill assessment and testing systems for basic skills and literacy, as well as other areas of instruction.

These criteria address "testing systems" rather than "tests" (e.g. instruments and questionnaires) because the procedures and institutional arrangements that determine the use of "tests" are critically important.

The ten criteria proposed in this chapter are: availability of technical documentation, applicability of test to multiple uses, pertinence of scores and measurements, validity and reliability, representative and comparative data, linkage to other tests and measurements, reporting and analysis, individual rights and privacy, administrative efficiency and viability, and procedures for review and test development.

Each of these ten criteria cover a number of secondary or detailed subpoints (Exhibit 2).

AVAILABILITY OF TECHNICAL DOCUMENTATION

All tests should provide technical documentation that is comprehensive, rigorous, accessible and understandable to users.⁸

All tests should provide technical documentation that is comprehensive, rigorous, accessible and understandable to users.

Technical documentation should cover issues of importance to the prospective uses of the test. In the case of skill assessment and testing systems, this documentation should provide information concerning all the criteria specified in this report.

Technical documentation also should be rigorous. It should meet professional psychometric and statistical standards. Wherever possible, it should be empirically based, and provide full disclosure of the data and methodologies used to produce the documentation. Data and information reported in the documentation should be updated regularly.

⁸ *Standards for Educational and Psychological Testing*. American Psychological Association, Washington D.C., May 1986, pages 35-38.

Exhibit 2
ELABORATION OF CRITERIA
FOR SKILL ASSESSMENT AND TESTING SYSTEMS

- (1) **Availability of Technical Documentation**
 - (a) Cover Issues or Criteria (See Below)
 - (b) Accessible and Understandable
 - (2) **Applicability of Test in Multiple Uses**
 - (a) Counseling and Feedback to Learners
 - (b) Placement and Certification
 - (c) Program Review, Evaluation and Planning
 - (3) **Pertinence of Scores and Range of Measurements**
 - (a) Type of Skills Assessed
 - (b) Level of Range of Skills
 - (c) Appropriate to Testing Groups
 - (d) Type of Measurement
 - (4) **Documentation of Validity and Reliability**
 - (a) Capacity to Measure Specified Skills and Knowledge
 - (b) Consistency of Measurement
 - (5) **Collection of Representative and Comparative Data**
 - (a) Capacity to Generalize Aggregate Results
 - (b) Capacity to Establish Norms and Comparison Groups
 - (c) Capacity to Measure Skill Improvement
 - (6) **Common Measurement of Skill Levels**
 - (7) **Capacity for Analysis and Reporting**
 - (a) Access to Learners and Other Test Users
 - (b) Nature and Relevance of Standard Reports
 - (c) Options for Analysis of Individuals and Subgroups
 - (8) **Safeguards to Protect Individual Rights and Privacy**
 - (a) Explanations to Test Taker and Other Impacted Groups
 - (b) Options for Review and Retesting
 - (c) Safeguards for Privacy
 - (9) **Cost Effectiveness and Administrative Efficiency**
 - (a) Cost and Efficiency of Operation
 - (b) Logistical Viability:
 - Planning administration
 - Test administration (test time, etc.)
 - Scoring and data entry
 - Data analysis and reporting
 - (10) **Procedures for Review and Test Development**
 - (a) Open Data for Review
 - (b) Third Party Evaluation
 - (c) Ongoing Research and Development
-

APPLICABILITY OF TEST TO MULTIPLE USES

Testing systems should be suitable for multiple uses, and these conditions should be specified. Adult skill assessment and testing systems should be usable for three purposes:

- (1) Counseling and Feedback to Learners
- (2) Placement and Certification
- (3) Program Review, Evaluation and Planning

Considerations concerning these three uses are discussed in this report.

The skills that assessment tests measure should be rigorously specified and pertinent to the skills that users wish to measure.

PERTINENCE OF SCORES AND RANGE OF MEASUREMENTS

The skills that assessment tests measure should be rigorously specified and pertinent to the skills that users wish to measure. This is necessary to ensure that test users understand what the test does and does not measure. It is also essential if users are to assess whether the test truly measures the skill attainment goals specified in program quality standards. In applying this criterion, prospective test users should consider five key issues:

- Types of Skills Assessed
- Level and Range of Skills Assessed
- Appropriateness to Testing Groups
- Types of Test Scores

These four issues are summarized below and elaborated in the full report.

Types of Skills Assessed

Test users should determine whether the measurements provided by the testing system match the learning goals of instructional programs. Unless test measures are matched with learning goals, information about student and program performance will be at best irrelevant and at worst inaccurate.

Level and Range of Skills Assessed

It is important to insure that the range of test measurements is appropriate to prospective uses of the test. A test that measures reading proficiency from basic skills to college level will generally be inappropriate to learners who are marginally literate. Further, wide ranges of measurement commonly lead to a lack of measurement precision.

Appropriateness to Testing Groups

Test questions, scoring scales and organization should be appropriate to the population tested. Most particularly, tests should be neutral and

unbiased to the range of ethnic, racial, age and gender groups that are expected to take the tests.

Types of Test Scores

Test measurements or scores should be pertinent and viable to the intended uses of tests. Just as distance can be measured with a variety of different scales (e.g. yards and meters), skills can be measured by a number of different scoring systems.

For the most part, the suitability of test scores will be determined by user frames of reference. Put differently, scores and measurements will be most meaningful and usable if they reflect the "real world" concern of test users rather than abstract numbers. For example, test scores may reflect whether test takers can read, perform the duties of a job, or possess skills equal to most adults.

Among the frame of reference issues of concern to users of adult skill assessment tests, three are immediately apparent:

- School Grade Level Reference
- Fixed Level Improvement Reference
- Criterion- or Norm-reference

In most cases, the selection of one frame of reference should not preclude use of other frames of reference. Test scores can be "crosswalked" to determine equivalent scores to other frames of reference, and scores can be analyzed in a variety of ways (e.g. performance on a skill level scale can be viewed in absolute terms or in terms of how the test score compares to those of other test takers). However, the primary scoring system of a test commonly reflects a focal frame of reference, and this frame of reference becomes the primary perspective that guides use of the test.

Validity and reliability are important criteria for evaluating tests. Validity is the extent to which a test measures what it claims to measure. Reliability is the test's ability to measure consistently.

DOCUMENTATION OF VALIDITY AND RELIABILITY

Validity and reliability are two of the most important criteria for evaluating tests.⁹ These two elements are interlinked:

- **Validity** is the extent to which a test measures what it claims to measure. For example, a test that claims to measure basic math will be invalid if it only measures capacity to conduct addition and division, or uses written questions that prevent respondents with literacy problems from demonstrating their math skills.¹⁰
- **Reliability** is the ability of the test to measure what it purports to measure consistently among respondents or over time. For example, a literacy test would have poor reliability if test scores

⁹ *Standards for Educational and Psychological Testing*, American Psychological Association, Washington D.C., May 1987, page 9.

¹⁰ Some social scientists divide validity into internal and external validity (see Donald T. Campbell and Julien C. Stanley, *Experimental and Quasi-Experimental Designs for Research*, Rand McNally College Publishing Company, Chicago, Illinois, pages 5-6). This section deals with "internal validity." "External validity," which is the extent to which measured results can be generalized or compared to larger groups, will be discussed in a later section dealing with the criteria of representative and comparable data.

varied greatly among respondents who have equal levels of literacy, or if a test that originally indicated high literacy levels revealed low literacy levels when taken by the same group three weeks later.

Both validity and reliability are determined not only by test design, but the procedures by which tests are administered. These two interconnected criteria are discussed below.

REPRESENTATIVE AND COMPARATIVE DATA

Test results are not used in a vacuum. They are commonly interpreted in reference to comparison groups, and increasingly used to analyze the status of tested populations. Test users should therefore consider the following three characteristics.

- Capacity to Generalize Aggregate Results
- Capacity to Generate Norms
- Capacity to Measure Improvement

The above three capacities generally depend more on testing procedures than test design. Each of these three capacities is elaborated in the following pages.

CAPACITY TO GENERALIZE AGGREGATE RESULTS

The capacity to generalize test results to populations targeted for testing is becoming increasingly important. The need for this capacity stems largely from pressures to evaluate the performance of social programs of standardized measures of their impacts on participants, and to assess the conditions of our general population. This is particularly true for education and training programs. Specifically, there is a growing need to measure the results of education in terms of whether students reach average societal competencies or show notable skill improvements. Similarly, there are needs to assess whether our total population or specific sub-populations have achieved the skills and knowledge that is required to function effectively in today's society.

Concern over program performance and the status of large populations adds an important new dimension to testing. In the past, testing and test results have focused on the individual. In the future, tests will continue to focus on the individual. However, they must also address the issue of whether test results can be aggregated to generalize the participants of programs or specified populations.

This growing need to generalize test results adds to the criterion of creating representative data. For the most part, this is a matter of test administration rather than test design.

There are two general approaches to collecting representative test data. The first is to test everyone in the population of interest. This is called a census. If a subgroup of the population is tested, the result is a sample. In many cases, taking a census is not feasible. As a result, the main challenge becomes getting a sample with representative data on test scores. In this

Concern over program performance and the status of large populations adds an important new dimension to testing. They must address the issue of whether test results can be aggregated to generalize the participants of programs or specified populations.

There are two approaches to collecting representative test data. The first is to test everyone in the population of interest. This is called a census. Taking a census is not always possible. As a result, the main challenge becomes getting a sample with representative data on test scores.

section the merits of a census and a sample will be discussed and the critical importance of representative data will be explained.

The primary tasks of sampling are to collect data from a representative portion of a population that is large enough to produce acceptable margins of error and allow analysis of subgroups of the sample.

Representative data hinges upon *random sampling*. Random sampling uses blind selection procedures where everyone in the target group has an equal chance of being selected for an interview. Therefore, every person in the target population to be studied will have a chance to be in the sample.

To illustrate the dangers of "non-random sampling," assume that a phone survey is being conducted, and that the target group is all students who graduated from a specific adult school. If data are collected from a "convenience sample" that saves phone toll charges by interviewing only those who are in the local call area, the sample will *not* be random. Maybe those who moved out of the area got better jobs because they were willing to move. The results of the interviews could lead you to conclude erroneously that their education did not help them get jobs. However, if each completer was assigned a number and a few of the numbers are blindly drawn "out of a hat" (or a computer), a random sample will be created to provide representative data.

It is also important to recognize that representativeness is far more important than the size of a sample. Tests should be administered to a small sample that is randomly selected to be representative rather than a large number that is not representative. This will yield better information than thousands of poorly collected responses:

Thus, test users should be keenly aware that the amount that scores collected from a randomly selected sample differs from the population of interest will decline as the sample size increases. Data from survey samples will always have some error. Data from samples will vary somewhat from the actual population. This is called *sampling error*.

The amount of "shake" or "error" in our sampled measurements is reduced as the number of observations increases. Imagine that you are sampling all the coin flippings in the universe. For this infinite number, you know that the average will be 50/50 for heads and tails. But this average will vary with the number of flips you make. If you take a "sample" of ten flips, the results could easily be seven heads and three tails. But if your coin flipping sample has 100 tosses, the chances are much greater that the results will be closer to the 50/50 distribution for the universe.

While it's always good to have large samples, there are points where the costs of added data collection do not result in significant reductions of sampling error. Thus, it is important to recognize that after a certain point, the reduction of error gained by increasing the size of random samples is not worth the added costs of data collection.

High response rates are a critically important principle of sampling. If persons are selected to be part of a representative sample, it is extremely important that test scores be collected from them.¹¹

¹¹ The *response rate* is found by dividing the number of completed interviews by the number of expected interviews. For example, if you drew up a list of 100 people to be interviewed and only managed to interview 60, your response rate would be 60 percent.

The primary goal of probability sampling is to collect a sample which is representative of the population of interest. As noted above, the key rules for such sampling are that every member of the population of interest must have an equal chance of being selected for the sample. It is equally important that sampled persons in the least do not select themselves out of the sample (i.e. not responding or refusing to respond). Thus, every effort should be made to collect data from all selected persons in the sample to avoid self-selection bias.

CAPACITY TO GENERATE NORMS THROUGH COMPARISON GROUPS

Measurement of skill improvement among learners is one of the most important roles of skill assessment testing. This requires the administration of pre/post tests to allow comparison of learner skill levels at the beginning of program participation, and at one or more points after learners start programs. It also requires an inventory of interchangeable test questions which can be used to create new tests about the same skills so that post-test scores reveal skill improvement rather than test similarity.

It is highly desirable to collect test data that allows flexibility comparing test scores by the characteristics of individuals and the programs they use. This analysis requires test data to be individual rather than batched program data.

It is also highly desirable to collect test data that allows flexibility comparing test scores by the characteristics of individuals and the programs they use. For example, it may be useful to see how the scores of minorities compare to those of non-minorities within different program settings. Or it may be desirable to isolate the individual characteristics of adult learners that are associated with success or failure to improve skill scores. These and other similar types of analysis require that test score data be coded and scored as individual rather than batched program data. Test users should consider the importance of the flexibility to analyze data when selecting tests, and seek tests and accompanying data storage and retrieval systems which allow recall of test scores by individual rather than group averages within programs.

COMMON MEASUREMENT OF SKILL LEVELS

It is extremely important that assessment tests provide a common "yardstick" for measuring skills and knowledge. Clearly, commonality *between* tests measuring literacy and parenting skill is not important. But commonality *among* tests measuring the same skills and knowledge is. From the standpoint of assessment and counseling, it is important that tests used by one program or locality be usable in another, and that tests administered at one point in time be comparable to those administered at a later point of time. From the standpoint of placement and certification, it is important that those who use tests to certify skills and make placement referrals are making decisions based on common measurements that are also understood and used by those relying on skill assessment tests as a criteria for employment or acceptance into instructional programs. From the standpoint of program planning and evaluation, it is important to have common measurements to assess learning outcomes within and among programs.

The most obvious approach to ensuring reliable common measurements is to use one test for skill assessment. This is not to say that other tests cannot be used, but that one test be required to provide "core" or a common linkage for the multiple applications of skill assessment tests.

From the standpoint of program planning and evaluation, it is important to have common measurements to assess learning outcomes within and among programs.

If one test cannot be used, the remaining option is to develop "crosswalks" that enable test users to translate the scores of different tests into common measurements. This option is generally viable and appropriate to a world with a significant plurality of testing systems. However, like the option of using one test, crosswalking test scores has limitations and shortcomings.

The reality is that skill assessment, whether with one or many tests, will never be perfectly accurate or without administration problems. Where possible, test users should select or develop one test that meets all needs. If this is not possible, test users should seek crosswalks with documented validity and reliability that allow the translation of test results with a common scoring system.

CAPACITY FOR ANALYSIS AND REPORTING

Capacity for test score reporting and analysis is an important consideration. Unfortunately, the design of many tests precludes the possibility of important types of analysis. In other cases, data storage and report production procedures do not allow or produce many needed reports.

Test users should consider three major issues dealing with the capacity for analysis and reporting:

- Nature and Relevance of Standard Reports
- Capacity to Analyze Individuals and Subgroups
- Access to Learners and Programs

It is critical that the standard reports produced by tests be relevant to user needs and that they provide technical documentation about the limitations of interpretation.

Test users should also consider the capacity to analyze test scores for individuals and specified subgroups. This capacity should not be taken for granted. Many tests only store and report scores for batched groups of test takers (e.g. classes or schools), and do not report individual scores. Conversely many testing systems have limited capacity to generate scores for specified groups, a shortcoming that greatly undermines their value for setting norms or program planning. To maximize the uses of tests, users should select or develop tests that store data on individuals, collect demographic data on each test taker, and link scores from tests taken by the same individual over time.

- Store Data on Individuals.
- Collection of Background and Demographic Data on Individuals.
- Unique Identification Number to Link Individual Scores Over Time.

The capacity to recall and analyze test scores for individuals and subgroups is fundamentally important to many applications of test data.

Access to test score data is also a major consideration for learners and programs. To the extent that it is economically possible and appropriate in terms of maintaining individual privacy, test data should be made readily available to users.

PROTECTION OF INDIVIDUAL RIGHTS AND PRIVACY

Skill assessment and testing systems should incorporate safeguards to guarantee maximum individual rights and privacy for test takers. Key considerations that should be addressed include:

- Explanations to Test Takers
- Options for Review and Retesting
- Safeguards for Privacy

These three considerations are discussed in the full report.

Skill assessment and testing systems should incorporate safeguards to guarantee maximum individual rights and privacy for test takers.

COST EFFICIENCY AND ADMINISTRATIVE EFFICIENCY

To have meaningful impact, tests must be usable in diverse "real life" circumstances. It is one thing to specify goals and standards for testing. It is another to implement a widespread skill assessment and testing system that serves intended purposes and meets acceptable technical standards. Two issues are particularly important:

- Cost Effectiveness
- Administrative Efficiency

If tests are to be used with the technical rigor needed to provide information worth using, they must be worth the cost. It is desirable for the costs of testing to be low. However, the real issue is "cost efficiency." Test users must ask whether the costs of using a test are greater or lesser than the costs of not using it.

Given acceptable costs, or a willingness to pay high costs to get acceptable test performance, test users must also consider the administrative viability of tests. It is possible that regardless of financial commitments, testing may be logistically unfeasible under some circumstances. Under some conditions, testing may be disruptive, require excessive time, or undercut learner morale. The administration of tests may be logistically infeasible under conditions of decentralized instruction or diverse course schedules. Test users should carefully consider these and other factors when selecting or developing tests.

TECHNICAL AUDITING AND TEST DEVELOPMENT

Tests should be evaluated not only for what they are but what they are becoming. Test users should select or develop tests that have institutionalized an open process of review and feedback, and that have long-term research and development plans. Three issues merit particular attention:

- Open Data for Review
- Third Party Evaluation
- Ongoing Research and Evaluation

To the extent possible within the constraints of protecting the personal privacy of test takers, testing data should be made available for review and evaluation. Potential users should have the option of assessing the quality of test data to ensure that the test performs in accord with test descriptions.

Test users should seek and place importance on third party evaluations of the test. Important conditions to look for in such evaluations include:

- Evaluators with Appropriate Professional Qualifications.
- Evaluators Who Are Independent and Have No Vested Interests.
- Recent and Regular Evaluations.
- Balanced and Empirical Evaluations.

Evaluations will vary in content and approach from test to test. However, it is important that tests have objective evaluations that address a criteria of performance that is relevant to test users.

Skill assessment and testing systems are more than a set of questions and procedures. They are ongoing efforts to measure skills in a changing world. Test users must select or develop tests that are not only good today, but that evidence ongoing professional activity to make improvements to meet the needs of tomorrow. Attention should be given to evidence of past and current research and development, formal recognition of shortcomings matched by efforts to make improvements, prioritized agendas for future research and development, and a record of responsiveness and cooperation with clients.

The ten criteria described above propose a descriptive checklist of elements that users might seek in skill assessment tests. Key themes from these interlinked criteria are that tests should be relevant to multiple users, that the quality of test the quality of test data must be acceptable or tests may cause more harm than good, and that tests should represent a systematic combination of procedures, applications and services that meet the needs of users. The next chapter formulates key elements of these ten criteria into recommendations that can be modified and prioritized for use in selecting or developing tests.

Evaluations will vary in content and approach from test to test. However, it is important that tests have objective evaluations that address criteria of performance relevant to test users.

Recommendations

This chapter presents proposals for the development of quality standards and performance measures for adult education. To better organize these proposals, they have been divided into two groups:

- (1) Priorities and Uses
- (2) Criteria for Skill Testing Systems

The proposals in the Criteria for Skill Assessment Tests focus on important considerations for the selection or development of a skill testing systems for adult learners.

The first two of the four proposals in the Priorities and Uses section are focused on the relationship between quality standards and performance measures and the priorities for developing both. The last two proposals in the Priorities and Uses section deal with the role of skill assessment and testing systems as a performance measure and the need for a criteria for the selection and development of tests.

The proposals in the Criteria for Skill Assessment Tests focus on important considerations for the selection or development of a skill testing systems for adult learners. This second group of proposals is based on the background discussion in Chapter 3.

PRIORITIES AND USES

The first four proposals are concerned with decisions that need to be made before skill assessment and testing systems can be developed or selected (Exhibit 3).

Exhibit 3 PROPOSALS DEALING WITH PRIORITIES AND USES

-
- (1) Develop Performance Measures in Response to Quality Standards
 - (2) Prioritize Development of Quality Standards and Performance Measures
 - (3) Use Skill Assessment Tests as the Primary Performance Measure
 - (4) Develop a Common Criteria for Evaluating all Skill Assessment and Testing Systems
-

The above four proposals establish the context for ongoing work on Quality Standards and Performance Measures.

CRITERIA FOR SKILL TESTING SYSTEMS

The remaining proposals deal with a general criteria for selecting or developing skill assessment tests for all areas of adult education and training. These proposals provide ten elaborations of Proposal 4 (Exhibit 4).

Exhibit 4
PROPOSED CRITERIA FOR EVALUATING FOR SKILL
ASSESSMENT AND TESTING SYSTEMS

- 4-1 Availability of Technical Documentation
 - 4-2 Applicability of Testing Systems to Multiple Uses
 - 4-3 Pertinence of Scores and Range of Measurements
 - 4-4 Documentation of Validity and Reliability
 - 4-5 Collection of Representative and Comparative Data
 - 4-6 Common Measures of Skill Levels
 - 4-7 Capacity for Reporting and Analysis
 - 4-8 Safeguards to Protect Individual Rights and Privacy
 - 4-9 Cost Effectiveness and Administrative Efficiency
 - 4-10 Ongoing Review and Test Development
-

It is unlikely that any existing test will fully meet all criteria. However, it is important to establish a comprehensive criteria to guide efforts to develop the best possible testing system.

It is unlikely that any existing test will fully meet all ten of these criteria. However, it is important to establish a comprehensive criteria to guide efforts to develop the best possible testing system.

These ten elaborative subproposals are defined and elaborated in the full report.

The ten subproposals above compose preliminary criteria for selecting or developing skill assessment and testing systems. These proposed criteria are distillations of elaborations provided in Chapter 3. Many criteria may need further elaboration, and additional criteria may be needed.

Future Action

This chapter provides a brief summary of the issues reviewed in the four prior chapters and a summary of the future actions in research and planning needed to implement the recommendations made in this report.

GENERAL REVIEW

Program quality standards are a necessary prerequisite for the establishment of performance measures. It is important that these standards define the minimum requirements, the normal expectations and the conditions of excellence to ensure that the adult education system meets the needs of stakeholder. These standards should document measures of performance that focus on the skills that students will need for work and life, and the skill achievement expected to result from participation in instructional programs.

Skill assessment tests are recommended as the primary mechanism for measuring student outcomes.

General criteria are presented for selecting or developing skill assessment and testing systems. It is proposed that first priority be given to developing program quality standards and measures that are designed to measure basic skills and literacy for adult learners. The proposals in the report are designed to facilitate the selection or development of skill assessment and testing systems that can be used in both traditional classroom and non-classroom settings.

Ultimately, the proposed criteria should be used for selection or development of skill assessment and testing systems for all adult education programs. Where possible, scores from these testing systems should be used to guide students, certify skill attainment, and to evaluate the quality of the programs.

DEVELOPMENTAL CONSIDERATIONS

When the proposals of this report have been reviewed, modified and approved via advisory committees and agency heads, attention must be directed to the issue of how to put them into action.

It is important to recognize the need to have consistency in the implementation of these proposals. It is also important to recognize that many of the proposals put forward in this report are interdependent. Failure to apply one recommendation may undermine the successful application of others. Because high quality testing and assessment requires careful testing, management and analysis, there is a need for consistent and systematic work to establish the testing system that will provide the necessary results for all stakeholders.

It is vital that the testing system be seen as legitimate at the local level. This suggests that the testing system must be developed and maintained through an ongoing process of state-local collaboration.

When looking to the future, it is necessary to suggest the types of actions that may flow from the proposals presented in this report. The first alternative might be to select the tests as they currently are used. The second alternative is to alter existing testing systems to ensure that they are applicable to the needs of the adult education system. A third and admittedly unlikely alternative, would be to develop new testing systems.

In all of these categories there is a need to ensure that there is close integration of quality standards and performance measures. Thus, there is a need to ensure that the two procedures for development have a close working arrangement in the future.

Another area that requires work and creative innovation is the use or development of non-test performance measures. Such non-test performance measures may include student retention rates, surveys of student satisfaction, and certifications from professionally acceptable educational institutions.

Ultimately, it is vital that the testing system be seen as legitimate at the local level. This suggests that the testing system must be developed and maintained through an ongoing process of state-local collaboration.

There should be an increase in the level of test assessment and design. This is especially true for the development of multiple use tests and for the use of tests in nontraditional areas. Finally, there is a real need for a high quality tracking design of those who leave the educational system early. Work must also be continued to develop more effective and representative tests, and more efficient management and proctoring in the field.

FUTURE ACTIONS

This report is the first step in a multi-stage initiative to develop consistent Quality Standards and Performance Measures for adult education.

This report is the first step in a multi-stage initiative to develop consistent Quality Standards and Performance Measures for adult education.

Future planning and implementation will be conducted under the guidance of the Subcommittee on Quality Standards and Performance Measures of the Interim Adult Education Steering Committee. This work will be subject to the review and approval of State Superintendent of Public Instruction Bill Honig, State Community College Chancellor David Mertes, and other participating agency heads.

Approved priorities and criteria from this paper will be used in conjunction with agency developed program quality standards to review and recommend one or more skill assessment and testing systems for appropriate authorized areas of adult education. It is proposed that first priority be given to developing quality standards, skill testing systems, and other non-test performance measures for areas of instruction dealing with basic skills and literacy.

Pacific Management and Research Associates

This paper has been prepared under Contract 7147 from the Adult Education Unit of the Youth, Alternative and Adult Educational Services Division of the California Department of Education. It has been prepared to summarize background information and outline possible policy decisions. The contents do not necessarily reflect the policies of the California Department of Education, Community Colleges, Interim Adult Education Steering Committee or other institutions.

PMRA

429 J Street, Suite 400
Sacramento, California 95814
(916) 448-9038

Chapter 1

Introduction and Overview

The most important question in education is "What does the student need to learn?" Quality Standards and Performance Measures need to address this issue.

This is the first of three papers dealing with the implementation of quality standards and performance measures for California's adult education system. These papers are being prepared in response to a recommendation of the 1988-89 *Strategic Plan for Adult Education*.¹

This first paper will focus on *skill assessment and testing systems* as a core performance measure. The term *systems* is more appropriate to this discussion than *tests*. It communicates a more holistic view that encompasses how tests are administered, how they are interpreted and how these interpretations are used.

This paper will propose generic criteria for developing these testing systems to meet a variety of needs.

The second paper will be a feasibility study. It will focus on the use of quality standards and generic criteria to guide development of skill assessment and testing systems. This second paper will make recommendations regarding the appropriateness of existing assessment procedures, instruments and systems. It will also address the need for supplementary or alternative performance measures.

The third paper will recommend a plan to implement quality standards and performance measures. This paper will present background information, proposed criteria and policy options for selecting or developing student performance measures.²

¹ *Adult Education for the 21st Century: Strategic Plan to Meet California's Long-Term Adult Education Needs* (1989 Edition), pages 64-65.

² *Author's Handbook for Working Papers: Instructions of Policy Option Papers, Feasibility Studies and Implementation Plans*, October 4, 1989, submitted to State Department of Education, prepared by Pacific Management and Research Associates.

This first paper has five chapters dealing with the following topics:

- (1) Introduction and Overview of Recommendation
- (2) Background on Student Skill Assessment and Testing Systems
- (3) Criteria for Evaluating Skill Assessment and Testing Systems
- (4) Recommendations
- (5) Future Actions

The first chapter will provide a background on the strategic recommendation for *Program Quality Standards and Performance Measures* and outline the orientation of this paper.

1.1

History and Summary of Recommendation

In 1988 an Adult Education Advisory Committee was created to provide policy guidance on the future of public school adult education in California. This committee produced a report titled *Adult Education for the 21st Century*.³

Recommendation 7 of this report calls for Program Quality Standards and Performance Measures. The exact text of this recommendation is provided below:

*Ultimately,
quality standards
and accountability
must be defined
as measurable
goals and
conditions.*

Quality standards are needed to specify a commonly defined and agreed upon criteria for establishing program certification to receive funding and license to operate as an educational institution. They are also needed as a basis for awarding incentive grants and bonus funds for excellence.

Ultimately, quality standards and accountability must be defined as measurable goals and conditions. If there are no standard measures of program success or failure, there is no common and acceptable way to document performance. Without such documentation, accountability is at best an arbitrary assessment based on claims rather than proof of performance.

³ *Adult Education Needs for a Changing State: Discussion Paper*, prepared for the Adult Education Unit, State Department of Education, by Pacific Management and Research Associates; *The California Adult Education System: Background Paper*, prepared for the Adult Education Unit, State Department of Education, by Pacific Management and Research Associates; *Adult Education for the 21st Century: Strategic Plan to Meet California's Long-Term Adult Education Needs*, (1989 Edition).

This plan proposes development of *Quality Standards and Performance Measures* to guide the development of programs and measure their success. Both these guidelines and measures are to be developed and adjusted through a broad-based participatory process by the proposed *Adult Education Institute for Research and Planning* in accord with procedures proposed for *Collaborative Planning*.

The *Program Quality Standards* are to be applicable to varied instructional conditions and methods (e.g. classroom settings, open-entry/open-exit, computer assisted instruction). They will provide minimum standards for certifying educational providers and provide guidelines for funding incentives.

The *Performance Measures* should provide ways of documenting success in applying quality standards. These performance measurements should assess both individual and educational provider performance.

Recommendation 7
QUALITY STANDARDS
AND PERFORMANCE MEASURES

Develop, with local participation, Program Quality Standards and Performance Measures that are appropriate to adult learning.

Program Quality Standards should:

- Applicability to varied methods of adult education (e.g. tutoring, class instruction, TV courses, etc.).
- Provide minimum standards for certifying programs for funding and legal status as educational providers.
- Establish guidelines for excellence and criteria for funding incentives (e.g. quality standards, quality indicators and quality frameworks).

Performance Measures should:

- Provide standard assessment of the educational performance of individuals and providers.
- Emphasize outcomes and skill improvement rather than inputs and processes.
- Measure performance for a variety of educational endeavors and approaches.

These measures should not necessarily replace existing standards and measures, but provide a common standard that can be used by all providers.

Individual measurements should address the following:

- Scores from standard or criterion-referenced tests of basic education (e.g. literacy, computation) and special topics (e.g. history, chemistry, typing).
- Measurement crosswalks that allow translation of standard scores to other tests and measures.
- Other criteria and documentation that can be used to certify skill levels (e.g. experience, products).
- Consistent measures to assess both the educational needs and achievements of individuals.

Educational provider performance measures should address the following:

- Aggregate student achievement tests.
- Amount of student skill improvement as measured by accurate and representative pre/post tests (with some attention to obtaining data on learners who leave before completing programs).
- Measures of educational inputs (e.g. facilities, staff).
- Productivity measured by average student achievement per dollar.
- Benefits to students after completing programs (e.g. employment, earnings, satisfaction as determined by followup surveys).

Ultimately, these measurements should produce data that can be stored and accessed through the proposed *Integrated Adult Education Data System*.

The intent is to establish a "baseline" or core set of measures that will provide a common means of viewing performance among many different types of programs.

It should be emphasized that we do not propose monolithic and inflexible guidelines and performance measures. The intent here is to establish a "baseline" or core set of measures that will provide a common means of viewing performance among many different types of programs. All participating educational providers will retain their discretion to develop and utilize their own measures as they deem necessary. Similarly, commonly agreed upon measures should be subject to open and regular review.

There are many established systems of quality standards and performance measures. These must be reviewed, screened and adopted where appropriate. Development of standards and measures might be prioritized by instructional area (e.g. ESL programs, ABE and GED programs, vocational programs, etc.).

The specific wording of Recommendation 7 will be used as a guideline for developing policy options for Quality Standards and Performance Measures.

1.2

Focusing the Recommendation

Research and planning on Recommendation 7 should be prioritized before effective action can be taken.

Research and planning on Recommendation 7 should be prioritized before effective action can be taken. This section will define quality standards and performance measures, suggest priorities for developing quality standards and performance measures, and identify programs and instructional areas where these standards and measures are most needed.

DEFINING STANDARDS AND MEASURES

First, it is necessary to specify and elaborate the meaning of quality standards and performance measures:

- Program quality standards specify the expectations of stakeholders for programs. These standards can define minimum requirements, normal expectations and conditions of excellence.
- Performance measures provide a way of documenting whether programs meet or surpass quality standards. They can focus on student outcomes (e.g. achievement test scores, program completions) or program inputs (e.g. student-teacher ratios, teacher certification, dollars spent per student).

Both quality standards and performance measures can have many manifestations.

FOCUSING ON LEARNER ACHIEVEMENT

A key theme of Recommendation 7 is that quality standards and performance measures should focus on student outcomes—most notably, demonstrated improvement and mastery of specified knowledge and skills.

This focus stemmed from a strong belief among the 1988–89 Adult Education Advisory Committee that the goals of instruction must be clearly specified and that outcomes be viewed in terms of what students actually learn.

Without disavowing other standards and measures, learner attainment of knowledge and skills is one of the common denominators that can be prescribed, measured and compared across all programs.

This theme does not suggest that program inputs (e.g. modes of instructions, facilities, teacher certification, dollars spent per student, curriculum content, etc.) and other measures of program performance are unimportant. To the contrary, program-related standards and measures are truly appropriate in specific ways to specific modes of instruction. However, the diversity of instructional approaches made available by educational technology (e.g. interactive computer-aided instruction and distance learning) and the multitude of existing providers increase the need for a common denominator. Without disavowing other standards and measures, learner attainment of knowledge and skills is one of the common denominators that can be prescribed, measured and compared across all programs.

While there are many ways to measure "student outcomes," deliberations of the 1988-89 Adult Education Advisory Committee focused on the instruments and procedures that compose learner assessment and skill testing systems.

PRIORITIZING BASIC SKILLS AND LITERACY

Adult Education covers a wide range of instructional programs. Adult Schools and the Community Colleges offer ten authorized areas of instruction:

- (1) Elementary Basic Skills (ABE)
- (2) High School Basic Skills (ABE)
- (3) English as a Second Language (ESL)
- (4) Citizenship
- (5) Substantially Handicapped
- (6) Vocational Education
- (7) Parent Education
- (8) Older Adults
- (9) Health and Safety
- (10) Home Economics

It is impossible to develop quality standards and performance measures for all these areas simultaneously. Priorities need to be specified to provide a manageable plan for developing them.⁴

⁴ Many of these areas of instruction lend themselves to common standards and performance measures. However, several areas will require unique ones. Areas such as substantially handicapped, parent education, and older adults may not lend themselves to "paper and pencil tests" and assessment. They may require different types of assessment and testing.

Recommendation 7 does not prioritize areas of instruction. However, student enrollment patterns coupled with social and economic needs clearly point to *adult literacy*, as the first and highest priority. Adult literacy includes basic skills such as reading, writing, verbal communication and numeracy.

Quality standards and performance measures for adult literacy must address the demands of a changing world. The definition of functional adult literacy is changing as literacy needs in the workplace increase and literacy requirements for access and use of new information technologies grows. Traditionally, reading, writing and math constituted most definitions of adult literacy. In the 1990's:

Recommendation 7 does not prioritize areas of instruction. However, student enrollment patterns coupled with social and economic needs clearly point to adult literacy, as the first and highest priority.

American Business now considers six other groups of basic skills to be essential for any kind of employment... These other skills are knowing how to learn, listening and oral communication, creative thinking and problem solving, personal management, group effectiveness, and leadership.⁵

Educational and skill requirements of U.S. Jobs are increasing—41 percent of the new jobs in coming years will have levels of proficiency in language, reasoning and mathematical skill compared with only 24 percent of existing jobs.⁶

California is becoming the most multi-cultural state in the nation. Along with the shift in demographics comes an expanding definition of adult literacy that encompasses communication skills of oral expression, listening comprehension, interpersonal communication skills of clarification, participation in group discussion, and team building.

This report will propose general criteria to guide the selection and development of skill testing and assessment systems. These criteria will be used initially in the area of adult literacy, but they can be applied to skill testing and assessment systems for other areas of instruction.

INTEGRATION WITH OTHER STRATEGIC GOALS

The philosophy of the 1988-89 Strategic Plan for Adult Education was to develop initiatives that are complimentary and mutually supportive. One of the most important goals of this plan is to improve the access of learners to programs through better information and a more integrated approach to assessment and counseling. As such, every reasonable effort should be made to develop quality standards and performance

⁵ Anthony Carnevale, *Workforce Basic*, U.S. Department of Labor, 1989.

⁶ William Johnson and Arnold Packer, *Preparing Workforce 2000*, Volume I, Number 1, U.S. Department of Labor, 1984.

measures that support the processes of assessment, counseling and guidance proposed for the Community Adult Education Information Services (Recommendation 3) and Integrated Adult Education Data System (Recommendation 8).

1.3

LINKING STANDARDS AND MEASURES

Another intent of Recommendation 7 is that performance measures be linked to quality standards. Specifically, quality standards should be established, then used to guide the development of performance measures. In short, we should have a way to measure whether we achieve what we intend to accomplish.

This will require coordination of Adult Education Institute (AEI) implementation research with a number of agency projects. Most notably, both the State Department of Education and the Community Colleges have ongoing projects to establish curriculum quality standards for programs:

- **Adult Curriculum Framework and Program Quality Task Forces.** The Adult Education Unit of the State Department of Education has organized field-based task forces to establish curriculum frameworks and quality standards for authorized areas of adult instruction. To date, task forces have been established for English as a Second Language (ESL) and Parent Education. Task forces will soon be assembled for Adult Basic Education (ABE) and Vocational Education. The goal of these task forces is to establish common quality standards across the state. Ultimately, these task forces are expected to produce standards that will guide development of performance measures in their areas of instruction.
- **Title 5 Course Standards.** The Chancellor's Office of the California Community Colleges has established basic standards for credit and noncredit courses. While these standards are not detailed, they provide some framework for developing performance measures.⁷

⁷ California Title 5 paragraph 55002, entitled *Standards and Criteria for Courses and Classes*, covers the issues of quality standards and measures for class performance of learners in the Community Colleges. The system has organized its course offerings into four broad categories: AA Degree Credit, Non-Degree Credit, Noncredit, and Community Service Courses. Adult education classes, which are noncredit courses, require less formal methods of evaluation than do the noncredit courses. In these courses, the course outline of record requires that scope and methods of instruction be listed, and that the methods for evaluating whether stated objectives in the course have been met must be specified. In those cases where the Community College is offering courses normally offered by the high school or unified school district, there must be mutual agreement between the two delivery systems about the courses to be offered. This suggests that some points of common quality standards might be developed between the two systems using this mechanism. In filing out the form for approval of a noncredit course, CCC-456, the agency must categorize its primary method of evaluation of meeting course objectives as one of the following (1) standardized instrument measuring student subjective opinion.

The in-process status of program quality standards need not prevent work on other parts of Recommendation 7. Therefore, this paper will specify general criteria for the selection or development of skill assessment and testing systems.

The in-process status of program quality standards need not prevent work on other parts of Recommendation 7. This paper will specify general criteria for the selection or development of skill assessment and testing systems.

The development of a criteria for skill assessment and testing systems instruments is far more than an academic formality. There are a number of basic skills tests used in ABE programs. There is also much discussion about which tests are best for specific programs and for different assessment purposes. A notable part of this discussion concerns the use of tests for diverse adult populations, different student goals, different instructional content, and different assessment purposes. Finally, there is considerable debate about the role of testing and the conditions that make tests valuable tools. The result has been a lack of agreement about which existing skill assessment and testing systems are appropriate for different purposes and how information from these systems should be used.

There is a need to clarify the attributes of skill assessment and testing systems that will meet the varied needs of learners, educators and taxpayers. This clarification will foster consensus on general criteria for selecting or developing assessment instruments and procedures. Together with program quality standards, these criteria would then be used to select or develop skills assessment systems for adult literacy and other instructional programs.

1.4

STEPS TO IMPLEMENT RECOMMENDATION

Over the next two years, the Adult Education Institute for Research and Planning (AEI) will work with the Interim Adult Education Steering Committee and other projects to accomplish four steps to develop quality standards and performance measures:

- (1) Establish General Criteria for Selecting or Developing Skill and Assessment and Testing Systems.
- (2) Develop Program Quality Standards for Areas of Adult Instruction, with Basic Skills and Literacy Receiving Top Priority.

(2) standardized instrument objectively measuring student knowledge, (3) student satisfaction with their educational experience, (4) competency based written and practical tests which demonstrate the students' ability to apply skills and concepts learned to minimum standards established by the instructor, (5) COPES.

- (3) Use Quality Standards and the Criteria to Select or Develop Skill Assessment and Testing Systems.
- (4) Use Quality Standards to Select/Develop Authentic or Applied Performance Measures, as well as Paper-Pencil Tests.

This paper will deal with the first of the previous four steps.

Chapter 2

Background on Skill Assessment and Testing Systems

Today's skill assessment and testing systems emerged in response to need for a standardized way to measure skill levels. Existing tests have greatly enhanced our ability to fairly and consistently assess the skill gains and needs of learners. However, new roles for testing and advances in information technology indicate that tests must continue to be enhanced to meet the needs of the future.

This chapter will present background information about skill assessment and test systems. It will cover five key topics:

- (1) Testing and Changing Social Needs
- (2) Testing Needs of Stakeholders
- (3) Overview of Skill Assessment and Testing Systems
- (4) Considerations and New Developments
- (5) Importance of Testing Standards

These topics provide background for the planning and development of skill assessment and test systems that will meet minimal testing standards while serving a variety of needs.

2.1

Testing in the Context of a Changing Society

There is nothing new about testing. In its most rudimentary form, testing began with the early foundations of civilization, when human communities first achieved a body of knowledge and skills that had to be passed on from generation to generation, or person to person. Those responsible for imparting knowledge found ways to assess whether learners had acquired the necessary skills. This took the form of tribal rites of passage to test endurance or hunting skills. It evolved into the day-to-day watchfulness of the master over an apprentice and the verbal drills of a medieval tutor and his pupil.

Testing systems evolved to achieve economies of scale, to permit comparison of individual results to a larger population group, establish measurable standards for certifying individuals for further education or employment, and to assess the performance of large groups or educational institutions.

Testing, as we commonly know it today, began with the introduction of orchestrated group learning such as classroom instruction. This mode of education created the need for a fair and consistent way of assessing the skills and knowledge of individuals learning in such groups. History shows the emergence of such systems in the monasteries of the European middle ages and the Confucian civil service examinations of ancient China. For most of us, the quintessential example of testing comes with the teacher developed test administered in one intolerable wave after another as we progressed through years of schooling. These tests evolved as a common artifact of modern society with the emergence of schools, particularly the democratization of education as a public undertaking.

The next step in the evolution of testing was the emergence of large scale standardized testing systems to determine the knowledge or skills of individuals across geographic regions and school districts. Testing systems evolved to achieve economies of scale, to permit comparison of individual results to a larger population group, establish measurable standards for certifying individuals for further education or employment, and to assess the performance of large groups or educational institutions. Emergence of these systems was also fueled by advances in data processing technology and the science of tests and measurements.

The trend toward large scale testing systems was readily apparent as a fixture of our educational systems by the 1950's. Examples of such systems include state mandated elementary and high school achievement tests such as the Iowa Achievement Test, California Test of Basic Skills (CTBS), and the more recent California Assessment Program (CAP). It also took the form of national standardized college entrance examinations such as the Student Aptitude Test (SAT) and the American College Test (ACT).

Standardized testing systems have not occurred without criticism and reservations. Many educators question whether large scale testing, and the importance being placed on it, creates a situation where instruction is focused on passing tests rather than well rounded education. Others have raised many technical questions about these tests: whether they truly measure educational attainment, whether they discriminate against ethnic and other population groups, and whether they invade the privacy and curtail the opportunity of students. Progress has been made on many of these problems, but much improvement is still needed. In a review of the strengths and weaknesses of standardized testing systems, the American Psychological Association concluded that all such tests are not always soundly developed or appropriate to the ways they are used, but that *"the proper use of well constructed and validated tests provides a better basis for making some important*

The basic purposes of testing has remained constant throughout history. One purpose of testing is to determine whether individuals or groups have achieved a desired level of skills or knowledge.

decisions about individuals and programs that would otherwise be available."¹

In broad overview, the basic purpose of testing has remained constant throughout history. One purpose of testing is to determine whether individuals or groups have achieved a desired level of skills or knowledge. We use tests to provide learners with feedback on how well they are doing and what they need to learn. We use tests as rite of passage to certify whether one is qualified to assume social responsibilities or pursue further education. We use tests increasingly to provide a fair and consistent way of determining whether individuals have achieved required skills.

These roles for testing will continue. However, it is likely that the uses of tests and testing systems may change and grow in response to new social and economic needs:

- **Testing to Monitor the Skills of Our Society.** As the skill requirements of modern "information societies" increase, it becomes apparent that we cannot support a competitive economy nor maintain a humane society unless we achieve higher levels of education and ensure that all members of our society have achieved a basic education. Large scale testing systems provide one way to monitor effectively the skills of our adult populations.
- **Testing to Provide Consistent, Fair and Flexible Certification.** As educational approaches diversify to meet the needs of our diverse populations and capitalize on the potentials of educational technology, a common denominator will be needed to certify whether individuals have attained required knowledge and skills.
- **Testing to Plan and Evaluate Education and Training Programs.** The essential result of any educational approaches and program is the skill attainment of participants. Test results for different educational programs, for groups within a program or for a program as a whole provide important information for teachers and administrators at the local, state and federal levels. Such test results are needed for effective program planning and evaluation.
- **Testing as a Learning Technique.** Feedback is a critical component of education. Interactive feedback as learning occurs has been proven to be an effective stimulate for learning. Advances in educational technology are rapidly paving the way to highly

¹ *Standards for Educational and Psychological Testing*, American Psychological Association, Washington, D.C., May 1986, page 1.

New social needs and dramatic advances in information technology suggest that significant attention be directed to the design and application of skill testing systems.

stimulating interactive learning systems. Interactive testing that both guides learners and documents skill attainments will provide this feedback.

New social needs and dramatic advances in information technology suggest that significant attention should be directed to the design and application of skill assessment and testing systems. These systems will become more important, increasing the need to fairly and accurately measure skills and knowledge in ways that are valid to diverse population groups. They are also likely to be important and highly visible features of our society, increasing the need to design them to function smoothly and efficiently. Finally they will be required to meet many needs, requiring that every effort be made to integrate testing systems so that they serve multiple objectives (Exhibit 2.2).

The roles of testing and testing systems in the future are likely to become increasingly important and comprehensive. However, these roles are likely to be viewed differently by groups impacted by testing.

2.2

What Do Stakeholders Want from Testing Systems?

The stakeholders of tomorrow's skill assessment and testing systems will be learners, educators and taxpayers. Each group has specific needs. At the same time, there are many common themes that suggest the need to design a testing system that serves all needs in a cost-effective fashion (Exhibit 2.1).

Exhibit 2.1 SKILL TESTING NEEDS OF STAKEHOLDERS

Needs of Adult Learners

- Assessment of Educational Needs for Placement
- A Learning Plan Based on Student Goals and Specific Skills Needed to Achieve That Goal
- Feedback on Learning Achievement
- Certification of Skill and Knowledge Attainment for Education and Employment

Needs of Educators:

- Planning Instruction with Feedback to Learner
- Educational Output Data to Evaluate and Guide Programs
- Fair Testing System
- Efficient and Cost-Effective Testing System

Needs of Taxpayers:

- Monitor Skill and Knowledge Attainment of Population
- Program Performance Indicators and Cost-Effectiveness
- Certification of Skills

Exhibit 2.2 INCREASING ROLES OF TESTING

The roles of testing and testing systems in the future are likely to become increasingly important and comprehensive.

Roles of Testing:

Stages of Testing:

	Observational Testing:	Paper and Pencil Tests for Small Groups:	Existing Standardized Testing Systems:	Future Standardized Testing Systems:
1. Rite of Passage, Demonstration of Skills and Knowledge to Gatekeepers				
2. Feedback and Guidance to Learner				
3. Consistent and Fair Certification of Skills for Small Groups				
4. Standardized Measures to Certify Skills and Knowledge for Large and Diverse Population Groups				
5. Consistent and Fair Certification of Skills for Large and Diverse Population Groups				
6. Analysis of Variables Impacting Learning				
7. Feedback to Instructors and Education Providers				
8. Output Data for Evaluation and Planning of Programs				
9. Data to Monitor Societal Skill and Knowledge Attainment				
10. Interactive Feedback to Support Learning				

Key:



Strong Role



Limited Role



No Role

WHAT DO ADULT LEARNERS WANT?

Students want to learn in order to obtain jobs and to improve the quality of their lives. Testing and assessment are needed to guide placement where students enter educational programs. Testing and assessment are also needed to give learners and instructors feedback on learning achievement and whether they are meeting their objectives. This feedback should support diagnosis of learning needs and development of individual learning plans. Information that summarizes the program performance is also needed to help learners select programs that will best meet their individual needs. Finally, testing and assessment provide a mechanism for certifying that learners have obtained skills and knowledge that they are prepared for jobs or further education.

WHAT DO EDUCATORS WANT?

Adult education instructors want to facilitate student learning. Adult educators recognize the value of tests and assessment as a means of helping students achieve their learning objectives. They want a skill assessment and testing system that is fair and efficient. To be considered fair, a skill assessment and testing system must provide equal opportunity for all learners to demonstrate competency without bias due to gender, language or citizenship. Tests and assessment must also be efficient. They must accomplish desired ends without undue cost or time-consuming administrative burdens. Finally, testing and assessment systems must evaluate both learners and programs with recognition that many students have enormous hurdles to overcome.

The bottom line for educational programs is whether participating individuals learn. Taxpayers want to know if programs enhance the skills of adult learners.

WHAT DO TAXPAYERS WANT?

Taxpayers want accountability—they want to know that they are receiving benefits for tax dollars spent. The bottom line for educational programs is whether participating individuals learn. Taxpayers want to know if programs enhance the skills of adult learners. They want to know which programs are the most effective and efficient in transmitting important skills to adult learners. Finally, taxpayers want to know if students achieve the minimum level of skills required for jobs. Skill assessment tests are therefore needed to evaluate the educational outcome of programs and to certify individual achievement.

2.3

An Overview Of Skill Testing Systems

This section provides an overview of skill testing for adults and examines the differences between centralized and decentralized approaches to testing.

There are a multitude of skill assessment and testing systems. Many of these are not appropriate for adults. This section will provide a context for examining assessment and testing systems. It provides an overview of skill testing for adults within educational settings and examines the differences between centralized and decentralized approaches to testing.

SKILL ASSESSMENT AND
TESTING SYSTEMS WITHIN
ADULT EDUCATIONAL SETTINGS

Skill assessment testing within adult educational settings is commonly administered through local jurisdictions to comply with state and federal requirements. For the most part, these testing systems are applied with limited control by state or federal agencies. This section provides an overview of such arrangements within California and other states.

Skill Assessment and
Testing Systems in California

Prior to 1960, assessment and testing in the adult education system was restricted to the use of teacher-designed tests and existing testing systems...there were few common testing requirements, and even fewer common testing measurements.

Skill assessment and testing systems within California evolved for many reasons. Key factors include the need for economies of scale in the development and use of tests, federal and state law mandating comparable test measurements, the need to link assessment of learning to movements for "competency-based education", and growing pressures to measure educational attainment and the performance of instructional programs

Prior to 1960, assessment and testing in the adult education system was restricted to the use of teacher-designed tests and existing testing systems. For the most part, there were few common testing requirements, and even fewer common testing measurements.

Starting in the 1960's, statutory requirements began to push adult educators toward more standardized approaches to testing. The federal *Adult Education Acts* of 1964 required that all programs using Section 309 funds provide skill assessment test score results for their participants. The result was an increase in testing and reported test scores, but little improvement in testing approaches or standardization of testing results. Local education providers collected test scores through a variety of testing instruments (most notably the

In 1974, a federally funded national study on adult basic skills galvanized movement toward a more standardized approach to testing.

TABE, ABLE and CAT) and forwarded these results to the State Department of Education to meet requirements. However, these data varied greatly in quality and were largely incomparable.

In 1974, a federally funded national study on adult basic skills galvanized movement toward a more standardized approach to testing. This study, *Adult Performance Levels* (APL), reported that one in five American adults had functional literacy deficiencies. The study focused attention on the meaning of functional and academic literacy and stimulated discussion about development of an assessment test to measure functional skills for adults.

The movement toward more systematic testing was also encouraged by the "competency based adult education" (CBAE) movement. This approach to instruction stressed a "performance-based process leading to demonstrated mastery of basic and life skills necessary for the individual to function proficiently in society."² As curriculum was developed for this approach it became apparent that the "demonstrated mastery of basic and life skills" aspect of the CBAE approach needed consistent measures of performance. This focused attention on the need for a basic skills assessment and testing system for adults.

Starting in 1980, a consortium for adult educators was funded by the State Department of Education to develop an assessment system that would measure adult basic skills. This consortium adopted the name California Adult Student Assessment System (CASAS), and after a time, CASAS was renamed to Comprehensive Adult Student Assessment System. By 1982, the first tests had been field tested and made available to meet both local testing needs and complying with federal test reporting requirements.

In 1982, California adopted policies that required the use of a competency based approach in adult education. These policies required programs using federal *Adult Education Act* funds to implement a competency based approach within three years to continue to receive these funds. These programs included adult basic education (ABE), English-as-a-Second Language (ESL), and Vocational ESL (VESL).

Skill assessment was a key component of this 1982 mandate. Agencies that received Section 306 (now Section 321) funds were required to administer CASAS to a representative sample of their students.³ Additionally, a

² *Evaluation Study Report: Competency Based Adult Education in California*, CBAE Staff Development Project, San Francisco State University, San Francisco and CASAS Project, San Diego Community College District, San Diego, March 1987. See also, National Center for Educational Statistics, 1982, page 80, as quoted in Judy Alamprese, *Investing in Change: Competency-Based Adult Education in California*, California State Department of Education, March 1987, ERIC ED 289 010, page 8.

³ Judy Alamprese, *Op. Cit.*, page 13

system of pre/post testing was applied to determine how well students learned skills between the beginning and end of courses. These pre/post tests were used to monitor student progress in courses dealing with Adult Basic Education (ABE), English-as-a-Second Language (ESL), and Vocational English-as-a-Second Language (VESL). This application of CASAS and other tests to placement and monitoring is being integrated into a vocational assessment and placement model developed by the State Department of Education to ensure that learners are placed in programs appropriate to their educational needs and that they are assisted in efforts to continue learning activities to reach personal goals.⁴

A 1987 evaluation of the impact of the 1982 requirement for "competency based education" revealed that there was an increase in the use of CASAS between 1982 to 1985. More students were being tested and more classes were participating even though the state only required that each agency test a representative sample. In 1984-85, there were 6,413 sets of paired data from 984 classes. In 1988-89, there were 12,353 sets for 1566 classes.

During the late 1980's, pressures grew to use skill test data as a means of measuring program performance. This created a need to generate test data that is representative of all students participating in programs.

During the late 1980's, pressures grew to use skill test data as a means of measuring program performance. This created a need to generate test data that is representative of all students participating in programs.

Starting in 1989-90, the State Department of Education initiated efforts to regulate the sampling of students taking the CASAS test within agencies receiving Section 321 funds. Those with fewer than 1,000 students are now required to test all of their students during a census period of two weeks. If they have 1,000 or more students they can take a sample. While this step recognizes the importance of improved sampling, it is only the first step toward collecting representative test score data.⁵

Use of CASAS has grown dramatically during the 1980's. CASAS has clearly become the leading skill assessment and testing system within California. However, it is important to recognize that it is not used universally. Given that California adult schools and community colleges provided basic skills instruction to approximately one million students in 1984-85, and that 12,969 of these students were tested at least once, the CASAS test was administered to only 1.3 percent of all applicable students.⁶ Further, many school districts and

⁴ State Department of Education, Sacramento, 1986.

⁵ Sampling procedures for education providers with more than 1,000 students are, as yet, not specified to ensure standardized administration of the test. This means that the measurement of performance may change partly because of the uniqueness of each site's testing situation. Thus when comparing one program to another, or one site to another, differences in reported scores may be as much a result of testing procedures as differences in the skills of tested groups.

⁶ The number of basic skills students was estimated by multiplying the total 1984-85 ADA for Adult Schools and Community Colleges by an estimated average number of nine students per ADA. There are an estimated nine students per ADA because most

community colleges use CASAS to meet statutory compliance requirements, but apply other assessment tests for purposes of counseling, placement and certification. Other assessment tests in use within California include:

- MDTP (Math Development Test Program)
- TABE (Test of Adult Basic Education)
- ABEL (Adult Basic Education Levels)
- ACT (American College Test)
- Locally Developed Tests
- College Board Developed Tests

CASAS has clearly become the flagship among adult skill assessment tests used in California. However, increasing demands for accountability and better assessment make it necessary to review the status of CASAS and other systems prior to moving forward in the development of quality standards and performance measures.

Skill Assessment and Testing Systems in Other States

*Other states
have the same
needs and
problems with
concerning skill
assessment and
testing systems.*

Other states have the same needs and problems with concerning skill assessment and testing systems. New York and New Jersey illustrate developments beyond California's boundaries.

New York state recently began to collect demographic information and to test all adult learners with Test of Adult Basic Education (TABE). All incoming adult students are given the TABE locator test to determine their general skill levels. Within 12 hours of instruction, they are given the regular TABE test at the level indicated by the TABE locator test. After approximately 100 hours of instruction they are given a followup test. The average change in scores is then calculated for the class. This followup is repeated before test results are sent to the state. New York requires use of the same test in all programs, and it requires that the system provide reports on gains in grade levels. Demographic information is analyzed to determine how learner backgrounds influence performance.⁷

New Jersey has an adult basic skills testing system which is similar to New York's incoming students are tested within 10

adult enrollment is part time. These calculations were based on data cited from Barry Stern, *The California Adult Education System: Background paper on the Response of Adult Education Institutions to the Needs of Californians*, Pacific Management and Research Associates, Sacramento, February 1989, pages 20 and 40.

⁷ Telephone interview with Dr. Julia Goodwin, Program Development Supervisor, State Department of Education, New York, January 10, 1990.

hours using either the TABE or Adult Basic Education Levels (ABLE) reading comprehension tests. The test is repeated after 50 hours, and a gain score is calculated. New Jersey, which has a newer system than New York, is experiencing administrative problems. In some cases learners are not given the locator test or they are not given the appropriate form of the test.⁸

The challenges of skill assessment testing within adult education settings are the result of multiple factors. First, American school systems have a traditional commitment to local control. This local autonomy makes it difficult to require use of a common test system and standardized testing procedures. Second, resources are scarce within adult education, and providers have several constraints concerning how much time and money they can commit to testing.

CENTRALIZED TESTING SYSTEMS

Many barriers to the development of comprehensive skill assessment and testing systems stem from the inability to enforce common test measures and standard testing procedures.

Many barriers to the development of comprehensive skill assessment and testing systems stem from the inability to enforce common test measures and standard testing procedures. This section explores the conditions necessary to establish common tests and testing procedures by examining the highly-regulated skill assessment system of the military.

The Department of Defense (DOD), with its unique opportunity to regulate the behavior of its "students", operates a highly-regulated testing system. Since civilian teachers do not have as much influence on their students as drill sergeants exercise over their recruits, this system would not be as successful in adult educational settings. However, the model does provide a good example of rigorous and centralized testing standards for the guidance, training and evaluation of adult populations. Some of these practices might be used to improve civilian testing and teaching systems.

The DOD uses the Armed Services Vocational Aptitude Battery (ASVAB)—a highly developed predictive measure of intellectual aptitudes for military occupations. This type of testing, which relies on the works of Galton and Binet, has been used to screen incoming recruits since World War I.⁹ This test, which includes the Armed Forces Qualification Tests (AFQT), takes about three hours to administer.

The DOD tests, trains and retests many times with the same instrument. Records have been maintained on individual

⁸ Telephone interview with Mr. Harry Van Houten, Manager of the Office of Program Development, Division of Adult Education, New Jersey State Department of Education, January 10, 1990.

⁹ Thomas G. Sticht, "Cognitive Science and Human Resources Management," *Advances in Reading/Language Research*, (Greenwich, Connecticut, JAI Press, 1982), page 5.

recruits for most of this century to predict performance in military occupations. As a result, the military has excellent comparative data.¹⁰ Testing of recruits and potential recruits has resulted in one of the largest sets of data on intellectual aptitude and performance found anywhere in the United States. The military can examine the impact of program changes on new recruits over time. At a given point in time they can compare the impact of various programs on recruits with similar background characteristics. Or the military can compare the impact of a program on groups of recruits with different background characteristics. Such systems of testing and analysis can provide high quality information for planning, programming, teaching and guidance.

During periods of high labor demand, the DOD's testing system allows them to train and use youth generally regarded as unemployable. In the Vietnam era, the DOD initiated Project 100,000 annually provided opportunities for up to 100,000 youth whose low aptitude test scores would normally have excluded them from the military service. Their testing system allowed the DOD to learn a great deal about the skill functioning and learning capabilities of these potentially "cast off" youth. This enabled the DOD was able to develop an effective means to train these recruits. As one study reported, "All told, the armed services appear to have repeatedly trained and used hundreds of thousands of youth deemed unemployable on the basis of aptitude test scores. Given an opportunity to prove themselves, and with support along the way, thousands of previously discarded youth used military service to break the cycle of poverty that had caught them and their parents."¹¹

While large, centralized systems such as the ASVAB may not be transportable to adult civilian populations. Some of the approaches used by such systems are worthy of further study.

The success of DOD's testing system is at least partially due to its centralized model of assessment, guidance, placement, and training. The virtues of this testing system have won the accolades of California Superintendent of Education, Bill Honig. "The advantages offered to students through the scheduling of ASVAB are indeed exceptional. I strongly recommend that the ASVAB be considered for use in all California high schools as a primary aptitude test for as many of our high school juniors and seniors as possible."¹²

While large, centralized systems such as the ASVAB may not be transportable to adult civilian populations. Some of the approaches used by such systems are worthy of further study. Advances in information technology are likely to enable civilian testing systems to provide many of the same functions

¹⁰ This reference is to the ASVAB14 (ASVAB A Brief Guide for Counselors and Educators, U.S. Government Printing Office, Washington, D.C., 1986).

¹¹ Thomas G. Sticht, *Cast off Youth Policy and Training Methods from the Military Experience*, Praeger Publishers, New York, 1987, page 182.

¹² Letter from Bill Honig, Superintendent of Public Instruction, California State Department of Education, to Secondary School Principals, May 29, 1986; Subject: Armed Services Vocational Aptitude Battery.

as those of the military without a highly-regulated system. Specifically, the use of computer-assisted interactive instruction, decentralized data storage through devices such as the proposed *EduCard™*, and efficiently designed sampling techniques hold the promise of providing more consistent and comprehensive testing.

2.4

Considerations and Developments

There are a number of considerations and developments to consider in the planning of quality standards and performance measures. These include new or changing theoretical perspectives, new technologies, and changing social and economic conditions.

THEORETICAL PERSPECTIVES

Two important theoretical perspectives to consider include andragogy and views of literacy.

Andragogy

Testing and instruction should reflect the needs of adults.

Testing and instruction should reflect the needs of adults. The instruction of adults has been named "andragogy". Andragogy is defined as "*the art and science of helping adults learn, in contrast to pedagogy which is the art and science of teaching children.*"¹³ The role of the learner is the key to understanding the difference between pedagogy and andragogy.¹⁴ In pedagogy the learner plays a dependent role. The teacher determines the subject matter, the teaching methods and the testing procedures. In andragogy, the students who are self-directed have some control over the subject matter, the teaching methods, and the testing procedures.

Andragogical principles have two implications for testing. First, testing should be self-paced and emphasize self-diagnosis. Second, test measurements and scales should be relevant to adult life.

Since adults learn differently than children, different education practices are needed according to Knowles. When

¹³ Malcolm S. Knowles, *The Modern Practice of Adult Education From Pedagogy to Andragogy*, (Chicago, Illinois: Association Press, Follett Publishing Company, 1980), pages 40-45.

¹⁴ Malcolm Knowles found that his principles of andragogy sometimes worked well with children, and he modified his definition of andragogy to be "simply another model of assumptions about learners to be used alongside the pedagogical model of assumptions, thereby providing two alternative models for testing out the assumptions as to their 'fit' with particular situations." (*Ibid*, page 43).

adults are tested, for example, greater emphasis should be placed on self-diagnosis.¹⁵ Knowles also suggested that adults may express different readiness to learn compared to children, and he suggested, "... learning programs should be organized around life-application categories and sequenced according to the learners' readiness to learn."¹⁶ Self-assessment at various stages of the learning process can be achieved through innovative techniques. For example, learners can experience job interviews in simulation exercises on computer systems with video disks.¹⁷ Other examples include computer assisted "tailored" testing, computer machine proctored examinations, and computer aided locator testing/evaluation.

Given that our highest priority for skill assessment test deals with basic skills and literacy, it is important to note the existence of different schools of thought about literacy.

The requirements of adult life raise issues as to whether test measurements should be reported in scales different from those used for children. Public discussions commonly describe skill requirements in terms of years of schooling (e.g. 8th grade equivalent). Such references provide a common reference point that everyone can understand. However, assessment of adults by such a scale may be demoralizing and possibly irrelevant.

Views of Literacy

Given that our highest priority for skill assessment test deals with basic skills and literacy, it is important to note the existence of different schools of thought about literacy. Definitions of "literacy" vary notably. In terms of testing, there are three schools of thought:

- (1) **Academic Definition.** One approach is to define literacy in terms of an academic or school grade equivalency. For example, people are considered to be literate if they function between the 4th and 8th grade level according to one definition.¹⁸
- (2) **Functional Life-Skills Definition.** A second approach defines people to be literate if they can function effectively within some social milieu. For example, functional literacy might be defined in terms of a person's ability to read a highway sign or newspaper article, or write a check. However, these tasks would have significantly different literacy demands.

¹⁵ *Ibid.*, page 47.

¹⁶ *Ibid.*, page 44.

¹⁷ Stich, "Cognitive Science and Human Resources Management," pages 312-313.

¹⁸ Ronald M. Cervero, "Is a Common Definition of Adult Literacy Possible?" *Adult Education Quarterly*, Vol 36, No. 1, Fall, 1985, page 50.

- (3) **Functional Job-Skills Definition.** A third approach defines literacy as the ability to function effectively in specific jobs. Since jobs require more focused or limited literacy skills, an acceptable job-literacy level may not be suitable to the literacy demands of life in general.

"Functional literacy is most usually thought of as the domain of literacy in the 'real world' outside of the school or academic setting."¹⁹ Often the definition of functional literacy is separated from the idea of academic literacy in that the former relates to being literate in the daily life activities of family, communities and work, while the latter often refers to reading, computational, and writing skills at a certain grade level (Exhibit 3).

Exhibit 2.3
ORIENTATION OF TESTS RELATIVE TO
THREE CRITERIA OF SOCIAL VALUE

Types of Tests:	Teaching Focus:	Rate of Change in Skills Acquisition:	Society's View of Skills
Test for General Literacy	Broad Academic	Slow	Comprehensive
Functional Adult Literacy	Broad Comprehensive	Moderate-High	Comprehensive
Test for Job Literacy	Narrow	High Partial	Job Focus

To develop enough tests to cover a myriad of job categories it would be necessary to use tests that relate directly to the job for which the person is being trained to test for functional or job related literacy. This suggests a difficult set of teaching, programming and job placement arrangements. Moreover, the tests themselves would have to be developed for a range of occupations and specified for each one to which the trainees are being oriented. This poses a difficult administrative problem.

There is evidence to suggest that students measured using this functional literacy orientation may demonstrate greater rates of accomplishment than students in the general literacy programs. Some research shows that adults learn job related reading skills more quickly than they do general reading skills, and that job related functional literacy training may have

¹⁹ Thomas G. Sticht, "Issues in Indexing Functional Literacy." A paper commissioned by the Study Group of the National Assessment of Student Achievement, Appendix B of *The Nation's Report Card*, February, 1987, ERIC ED279 700, pages 7-8.

potentially more rapid gains in literacy skills development for adults than does general literacy training.²⁰ These different definitions of adult literacy, and how to achieve it, are significant to the development of accountability in adult basic education programs.

Before deciding whether to measure functional literacy, job literacy or general literacy, adult educators will want to consider the learner's goals, the rate at which literacy can be achieved, the value of the skills acquired, and the expected public support.

Teaching and testing general literacy is likely to be more widely supported in the society, but learners are more likely to be able to attain functional reading and writing skills that lead directly to jobs.²¹

NEW KNOWLEDGE AND TECHNOLOGY

Newer more dynamic software developments give students greater control over self-pacing, editing, and data input. Such systems can be helpful to more closely integrate the adult learner with the subject matter because the learner is actually conducting self-paced instruction and assessments.

New knowledge and technology are likely and to change our expectations and options for adult assessment testing. Key issues to be considered include computer aided testing, new delivery systems, new measurement technology, and item-response theory.

Computer-Aided Instruction and Testing

Skill and drill computer programs are no better than traditional student workbooks according to Wangberg.²² However, software that is interactive and supportive offers several advantages over the older style skill and drill programs. Newer more dynamic software developments give students greater control over self-pacing, editing, and data input. The result is an andragogical approach especially suitable to adult education. Such systems can be helpful to more closely integrate the adult learner with the subject matter because the learner is actually conducting self-paced instruction and assessments.

Skill and drill programs are much like workbooks that require students to "fill in the blanks" or "match the terms" and compare their answers against those in the workbook or on the screen. It is a passive process perfectly suited for a pedagogical approach. Newer computer programs, described in research reported by Wangberg and others, are more suitable

²⁰ Thomas G. Sticht, et. al., *Cast-Off Youth Policy and Training Methods from the Military Experience*. (Praeger, New York, 1987), pages 132-134.

²¹ Thomas G. Sticht, "Evaluation of the Reading Potential Concept for Marginally Literate Adults," Alexandria, Virginia, Human Resources Research Organization, February 1982, ERIC ED217 168, pages 21-24.

²² Elaine Wangberg, "The Promise of Technology in the Reading/Writing Curriculum," *California Technology Project Quarterly*, Vol 1, No. 1, California State University, Chico, 1989, page 1.

for adults who want to learn literacy. First, students using a microcomputer with Wangberg's software are given a list of topics. After choosing a topic, the computer prompts the student to write his or her thoughts about the subject. Further prompts help the student develop the topic. If the student is having difficulty responding to the computer, an audio cassette will provide oral instructions. Later the student and teacher will further edit the lesson. In subsequent sessions the students can recall material that they have been working on for further development. By comparing the work they did in earlier lessons with their current writings, the students, with the help of their teacher, can engage in self-assessment. Such "Whole Language" programs may hold some promise for the development of written word assessment methods.²³

Since education is, in reality, a lifelong process, traditional educational institutions and teachers are no longer the only source of this service. These new approaches require that testing and measuring student performance be flexible enough to adapt to these varying circumstances, yet be scientific enough to provide legitimate bases for policy decisions.

The computer is especially suitable for adults who have irregular attendance since it allows them to review and pick up where they left off. No matter how many times the lesson needs to be repeated, the computer never loses patience.

The computer is well-received according to Wangberg who says that it is perceived as being more appropriate and more relevant for adults than a workbook. Perhaps this is because adults remember workbooks from their childhood days in school while they associate computers with the work place. Thus the computer may motivate students to stay with it longer than other learning activities according to Wangberg.²⁴

New Delivery Systems

Since education is, in reality, a lifelong process, traditional educational institutions and teachers are no longer the only source of this service. For example, people can enroll in continuing education classes through a local college that allow them to take their classes on television in their own homes. Or they can take basic job training classes on the local college campus funded by a federal program such as the Job Training Partnership ACT (JTPA). In some cases students may gather at a local business, such as the Holiday Inn, which has a satellite downlink for receiving classes produced at a distant site. These new approaches require that testing and measuring student performance be flexible enough to adapt to these varying circumstances, yet be scientific enough to provide legitimate bases for policy decisions. For example, a maximum ratio of 30 students per faculty member may be inefficient for televised classes, but testing may present problems at multiple remote sites.

²³ Elaine Wangberg and Mary K. Reuten, "Whole Language Approaches for Developing and Evaluating Basic Writing Ability," *Lifelong Learning*, June 1986, page 14.

²⁴ *Ibid.*, page 13.

New Measurement Technology

The old view of literacy resulted in a classification of the population into three subgroups: the literate, the illiterate, and the functionally illiterate. One new view, recognizes three different types of literacy that are measured on scales of prose literacy, document literacy and quantitative literacy.

- Prose literacy is the knowledge and skills needed to understand and use information from texts that include editorials, news stories, poems, and the like is defined by the National Assessment of Educational Progress (NAEP) test bank as a central feature of day by day literacy. Such ideas of literacy fit in closely with the notions of general literacy, and the research NAEP has done to estimate the number of persons in the society who are literate is well worth applauding.²⁵
- Document literacy is the knowledge and skills required to locate and use information contained in job applications or payroll forms, bus schedules, maps, tables, indexes, and so forth is another category of literacy that makes sense for daily life literacy concerns.²⁶
- Quantitative literacy is the knowledge and skills needed to apply arithmetic operations, either alone or sequentially, that are embedded in printed materials. Such daily operations as balancing a checkbook, figuring out a tip, completing an order form, or determining the amount of interest from a loan advertisement are considered to be normal activities a literate adult should be able to do on a daily basis.²⁷

There is no standard point on the scales for any of these three types of literacy that divides the literate from the illiterate, rather the NAEP measures literacy from low to high. In their sample of young adults (21 to 25 years old), Kirsch and Jungeblut found that an "overwhelming majority" could perform tasks on the lower end of the scales, but only a small percentage could perform the more complex tasks on the upper end of the scales.²⁸

²⁵ Irwin S. Kirsch and Ann Jungeblut, *Literacy: Profiles of America's Young Adults*, National Association of Educational Progress, Educational Testing Service, Report No. 16-PL-02, Princeton, New Jersey, 1986, page 4.

²⁶ *Ibid.*, page 4.

²⁷ *Ibid.*, page 4.

²⁸ *Ibid.*, page 4.

Item Response Theory (IRT), allows tests to be "tailored" to the appropriate level of difficulty for the student being tested. In the near future it seems that more sophisticated "tailored testing" procedures using this technology might be developed.

Kirsch and Jungeblut advocate measuring an individual's performance on each of the three scales with simulation tasks (e.g. writing a letter to explain an error in billing and giving an oral interpretation of a newspaper story). They argue that this approach is much better than multiple choice exams because it provides the opportunity to observe behavior under conditions that are more realistic. While this approach is commendable, it is also quite difficult to scientifically measure since it is a formidable task to carry out such program tasks across large numbers of students.

Item Response Theory

This important new development in testing and measurement, Item Response Theory (IRT), allows tests to be "tailored" to the appropriate level of difficulty for the student being tested. IRT is a statistical method that predicts examinee performance by using a known characteristic or trait of the person. For example, a provider might use the number of years of formal education to predict the general level of learner's performance on a prose literacy scale.²⁹ Some models use more than one characteristic to predict test performance. For example, a learner's age, I.Q. and years of formal schooling might be used to predict the general level of performance. Rasch scaling, a model of scaling within the framework of IRT, can be used to estimate the probability that a person located on an attribute scale will answer a question correctly. With this type of information it may be possible to calculate probabilities of correct responses prior to giving the test. In the near future it seems that more sophisticated "tailored testing" procedures using this technology might be developed.³⁰

The Rasch model can be used to develop comparative test scoring. It is possible to calculate difficulty measures for each item in a test bank. This information across several sets of tests might provide crosswalks that allow the provider to predict Jane's score on other tests. Further, it is argued that these procedures can produce ability grouping tests. These more powerful tests of persons grouped into levels might provide better information about change of learners resulting from teaching activities.³¹

CHANGING SOCIAL AND ECONOMIC NEEDS

Ongoing social and economic changes will also impact the need for adult skill assessment and testing systems. Two key trends to consider include the skill gap and demographic changes.

²⁹ Ronald K. Hambleton and Hariharan Swaminathan, *Item Response Theory*, Kluwer-Nijhoff Publishing, Boston, 1985.

The Skill Gap and Importance of Testing

A growing gap between the skill requirements of today's world and the skills of our population indicate that we must measure more effectively the skills of our society and provide feedback to individuals as to whether they have the skills needed to lead productive and rewarding lives.

A growing gap between the skill requirements of today's world and the skills of our population indicate that we must measure more effectively the skills of our society and provide feedback to individuals as to whether they have the skills needed to lead productive and rewarding lives.

By 2000, the skill requirements for a job will be equivalent to 13.5 years as compared to 12.8 years in 1984.³² About 41 percent of the jobs will require higher skill levels while only 27 percent will require lower skill levels in 2000. Although the average skill level of Californians is increasing, it is not rising as rapidly as job requirements. Almost 80 percent of the people who will be in the labor force in 2000 are in today's labor force. Their median education is only 12.5—one year less than the average required for 2000. This gap is more apparent among some subgroups of population than others. Although Whites, Asians and Blacks have on the average 12 to 13 years of education, Hispanics average only 9 to 9.6 years of education.³³

This gap between the present level of education and that required to fill available jobs fits andragogical definitions of educational need. "An educational need... is something people ought to learn for their own good, for the good of an organization, or for the good of society. It is the gap between their present level of competencies and a higher level required for effective performance as defined by themselves, their organization, or society."³⁴

Today's educational need, or skill gap, emphasizes the importance of providing effective learning opportunities for adults and developing reliable ways of measuring whether individuals and our society are attaining the skills and knowledge that are needed.³⁵

pages 8—9.

³⁰ *Ibid.*, pages 13-14 and 296-301. For a full exposition of the Rasch method and some of its mathematics see: David Andrich, *Rasch Models for Measurement*, Sage Publications, Newbury Park, California, 1988.

³¹ Jan-Eric Gustafsson, *An Introduction to Rasch's Measurement Model*, ERIC/TM Report 79, Educational Testing Service, Princeton, New Jersey, 1980, pages 32-39.

³² William Johnston and Arnold Packer, *Workforce 2000: Workers for the 21st Century*, Hudson Institute, Indianapolis, 1987, pages 96-101, as quoted in Fred Best, *Adult Education for the 21st Century: Strategic Plan to Meet California's Long-Term Adult Education Needs*, California State Department of Education, Adult Education Unit, May 1989, page 11.

³³ California State Department of Education, Adult Education Unit, *Adult Education for the 21st Century: Strategic Plan to Meet California's Long-Term Adult Education Needs*, May 1989, page 13.

³⁴ Knowles, *Op. Cit.*, page 88.

³⁵ *Ibid.*, page 19.

Demographic Changes

The increasing diversity of California's population indicates that testing systems of all types must be applicable to persons from different backgrounds.

The increasing diversity of California's population indicates that testing systems of all types must be applicable to persons from different backgrounds.³⁶ For example, tests must be culturally neutral to measure the competencies of Hispanics as well as Anglos. Using computer based testing in some classes may alter the performance of some examinees by creating excessive test anxiety. Some students from societies without such technology may find the computer too much to manage while being tested too. ESL classes may be particularly susceptible to this problem. As a result, ESL students may require different literacy programs and different testing systems according to Sticht.³⁷ Tests must also be as easily understood by senior citizens who may have difficulty seeing small print or hearing at normal voice levels.

2.5

Forging Skill Assessment Tests for Tomorrow

The changing nature of society and education have caused a major expansion in the roles of testing and paved the way for new and greater roles for testing to meet the needs of the future.

Over the last several decades testing has evolved from limited paper-and-pencil examinations developed and used by classroom teachers to large scale testing systems designed to provide standardized measurable skills and other attributes for large and diverse populations. Tremendous advances have been made to increase the reliability, relevance and efficiency of such testing systems. Nonetheless, virtually all these testing systems have shortcomings that require further development and refinement.

In coming years, the roles of testing in our society are likely to grow and diversify. There will be a need to better monitor the skill level of our population, document the learning output of our educational systems, and provide better and fairer assessment and certification of individual skills. These needs will be matched with new technologies and techniques that will enable us to expand testing service by better integrating them into the fabric of our educational systems. Testing is a powerful tool for assisting individuals and society understand where we are and what we must do to improve learning, and ultimately to reinforce the actual learning process. These goals

³⁶ *Adult Education for the 21st Century, Op. Cit.*, page 20.

³⁷ Thomas G. Sticht, "Issues in Indexing Functional Literacy Development," paper presented at the National Adult Literacy Conference, Washington D.C., January 1984, ERIC ED 240 300, page 7.

can be achieved. However, efforts must be based on a common agreement about our testing needs, understanding of minimal standards for testing, and commitment to marshal technology and resources to make testing a tool to enhance hope and opportunity within a context of accountability for learning.

Chapter 3

Standards for Skill Assessment and Testing Systems

It is important that skill assessment and testing systems meet minimal standards. This chapter outlines criteria to be used as general standards for the selection or development of skill assessment and testing.

This chapter will discuss ten standards, or criteria, for evaluating skill assessment and testing systems. These criteria are intended to serve as general standards for the selection or development of skill assessment and testing systems for basic skills and literacy, as well as other areas of instruction.

This report addresses "testing systems" rather than "tests" (e.g. instruments and questionnaires) because the procedures and institutional arrangements that determine the use of "tests" are critically important.

The ten criteria proposed in this chapter are: availability of technical documentation, applicability of test to multiple uses, pertinence of scores and measurements, validity and reliability, representative and comparative data, linkage to other tests and measurements, reporting and analysis, individual rights and privacy, administrative efficiency and viability, and procedures for review and test development (Exhibit 3.1).

Exhibit 3.1 CRITERIA FOR SKILL TESTING SYSTEMS

- (1) Availability of Technical Documentation
 - (2) Applicability of Test to Multiple Uses
 - (3) Pertinence of Scores and Range of Measurements
 - (4) Documentation of Validity and Reliability
 - (5) Collection of Representative and Comparative Data
 - (6) Common Measurement of Skill Levels
 - (7) Capacity for Reporting and Analysis
 - (8) Safeguards to Protect Individual Rights and Privacy
 - (9) Cost Efficiency and Administrative Viability
 - (10) Procedures for Review and Test Development
-

Each of these ten criteria cover a number of secondary or detailed subpoints (Exhibit 3.2).

This chapter will briefly explain how to interpret each of the criteria and why each is important to testing systems that measure the skills and knowledge of adult learners. Specific recommendations on how to apply these criteria will be presented in Chapter 4.

3.1

Availability of Technical Documentation

All tests should provide technical documentation that is comprehensive, rigorous, accessible and understandable to users.¹

All tests should provide technical documentation that is comprehensive, rigorous, accessible and understandable to users.

Technical documentation should cover issues of importance to the prospective uses of the test. In the case of skill assessment and testing systems, this documentation should provide information concerning all the criteria specified in this report.

Technical documentation also should be rigorous. It should meet professional psychometric and statistical standards. Wherever possible, it should be empirically based, and provide full disclosure of the data and methodologies used to produce the documentation. Data and information reported in the documentation should and updated regularly.

Finally, technical documentation should be accessible and understandable to test users. Summary documentation suitable for review by lay persons should be readily available, and detailed documentation should be available to technical readers. Where possible, this documentation should be released to third party reviewers and published in technical reports and journals (with comparisons to similar tests). Procedures should be established to ensure that test providers respond to reasonable requests for information from users.

¹ *Standards for Educational and Psychological Testing*, American Psychological Association, Washington D.C., May 1986, pages 35-38.

Exhibit 3.2
ELABORATION OF CRITERIA FOR SKILL
ASSESSMENT AND TESTING SYSTEMS

- (1) **Availability of Technical Documentation**
 - (a) Cover Issues or Criteria (See Below)
 - (b) Accessible and Understandable
 - (2) **Applicability of Test in Multiple Uses**
 - (a) Counseling and Feedback to Learners
 - (b) Placement and Certification
 - (c) Program Review, Evaluation and Planning
 - (3) **Pertinence of Scores and Range of Measurements**
 - (a) Type of Skills Assessed
 - (b) Level of Range of Skills
 - (c) Appropriate to Testing Groups
 - (d) Type of Measurement
 - (4) **Documentation of Validity and Reliability**
 - (a) Capacity to Measure Specified Skills and Knowledge
 - (b) Consistency of Measurement
 - (5) **Collection of Representative and Comparative Data**
 - (a) Capacity to Generalize Aggregate Results
 - (b) Capacity to Establish Norms and Comparison Groups
 - (c) Capacity to Measure Skill Improvement
 - (6) **Common Measurement of Skill Levels**
 - (7) **Capacity for Analysis and Reporting**
 - (a) Access to Learners and Other Test Users
 - (b) Nature and Relevance of Standard Reports
 - (c) Options for Analysis of Individuals and Subgroups
 - (8) **Safeguards to Protect Individual Rights and Privacy**
 - (a) Explanations to Test Taker and Other Impacted Groups
 - (b) Options for Review and Retesting
 - (c) Safeguards for Privacy
 - (9) **Cost Efficiency and Administrative Viability**
 - (a) Cost and Efficiency of Operation
 - (b) Logistical Viability:
 - ~ Planning administration
 - ~ Test administration (test time, etc.)
 - ~ Scoring and data entry
 - ~ Data analysis and reporting
 - (10) **Procedures for Review and Test Development**
 - (a) Open Data for Review
 - (b) Third Party Evaluation
 - (c) Ongoing Research and Development
-

3.2

Applicability of Test to Multiple Uses

Testing systems should be suitable for multiple uses, and these conditions should be specified. Adult skill assessment and testing systems should be usable for three purposes:

- (1) Counseling and Feedback to Learners
- (2) Placement and Certification
- (3) Program Review, Evaluation and Planning

Considerations concerning these three uses are discussed below.

**COUNSELING AND FEEDBACK
TO LEARNERS**

Testing systems should be suitable for multiple uses, and the conditions should be specified. Adult skill assessment and testing systems should be usable for three purposes: (1) counseling and feedback to learners; (2) placement and certification; and (3) program review, evaluation and planning.

Tests can be used to monitor and evaluate students. It is important that teachers have access to tests that will help in the placement, guidance, evaluation, and monitoring of persons through the learning process. Feedback is important to the learning process. Some say that it may be more important than other factors determining student success.² Standardized testing systems can provide students, teachers and counselors with information about the skill and knowledge levels of learners relative to other learners and the requirements of society. Such test scores should be used to reinforce learning achievement and make students aware of their learning needs.

**PLACEMENT
AND CERTIFICATION**

Skill assessment tests are also used to sort and screen learners for placement into instructional programs and to certify that learners have mastered certain skills. Tests should augment decisions made by counselors or entrance review staff. For example, the Scholastic Aptitude Test (SAT) is used to determine whether students should be admitted to many colleges and universities. In other cases, test results can be used for certification. For example, passage of General Education Development Test (GED) certifies high school equivalent skills in specified areas, and passage of the Bar Examination certifies attorneys to practice law.

² Thomas Boyson, *Examining State Schools Testing*, *The Sacramento Bee*, May 22, 1990, page B5.

PROGRAM REVIEW, EVALUATION AND PLANNING

A third application for testing is to evaluate how well education and training programs help students to learn. The most general case requires that learners be tested as they enter a program and then as they leave. The first score is subtracted from the second and it is assumed that the difference is explained by the teaching and learning process. Of course, all sorts of factors can influence the changes in behaviors measured between the pre and post scores. For example, the literacy gain may be overstated if learners score better on the second test because they have learned how to take tests. Or students may simply increase skills and knowledge over time by life experiences that are independent of instructional programs. Despite those problems, testing and analytical procedures can be developed, which make testing systems one of the most important tools for evaluating the performance of instructional programs.

It is highly desirable to accomplish many objectives with one test.

It is highly desirable to accomplish many objectives with one test. Under most circumstances this will reduce costs, increase communication and coordination through common measurements, and reduce administrative problems. However, some circumstances may require the use of several cross-referenced tests. There are numerous possibilities and considerations. For example, a single test could be used by all agencies for program evaluation and monitoring, while each agency is allowed to decide which test or tests to use for student placement, guidance, evaluation, monitoring and self-evaluation. If multiple tests are used, it is highly desirable that they be cross-referenced to allow comparison of scores—an issue that will be discussed later.

3.3

Pertinence of Scores and Range of Measurements

The skills that assessment tests measure should be rigorously specified and pertinent to the skills that users wish to measure. This is necessary to ensure that test users understand what the test does and does not measure. It is also essential if users are to assess whether the test truly measures the skill attainment goals specified in program quality standards. In applying this criterion, prospective test users should consider five key issues:

- Types of Skills Assessed
- Level and Range of Skills Assessed

- Appropriateness to Testing Groups
- Types of Test Scores

These four issues will be discussed in the following pages of this section.

TYPES OF SKILLS ASSESSED

Given that skill achievement goals are clearly designated by program quality standards, prospective test users should determine whether the measurements provided by the testing system match the learning goals of instructional programs. Unless test measures are matched with learning goals, information about student and program performance will be at best irrelevant and at worst inaccurate. Thus, deciding on appropriate measures of learner skills will be a central part of documenting achievement toward quality standards.

Given that skill achievement goals are clearly designated by program quality standards, prospective test users should determine whether the measurements provided by the testing system match the learning goals of instructional programs.

Manuals and documentation about the skill assessment test should clearly state the skills that the test is intended to measure. These definitions should be precise and detailed. For example, in the case of literacy there are many different choices for measuring skills. Some include basic academic skills, life skills and skills concerning specific topics (e.g. job skills or parenting).

Test users need to examine the documentation provided by the test supplier to determine if the skill measurement is well defined. The first consideration should be whether the test covers the skills that learners need and that instructional programs teach. The level of detail needs to be sufficient to leave no doubt in the test user's mind about the topics that are covered.

Sometimes tests are actually measuring accumulated knowledge, or prior experience, rather than skills taught in the class. The test user should be aware of what the test is expected to measure in order to avoid using resources inefficiently and appropriately.

LEVEL AND RANGE OF SKILLS ASSESSED

It is important to insure that the range of test measurements is appropriate to prospective uses of the test.

The scope of a test's measurement is important because it is an indication of the amount of knowledge or behavior the test is thought to measure. However, there are tradeoffs between

the breadth of test measurement and the precision of the measured behavior.

The scope of a test's measurement is important because it is an indication of the amount of knowledge or behavior the test is thought to measure.

Although narrow-scoped tests limit the knowledge and behaviors that are being assessed, they provide more precise measurement and more detail. Teachers might be more interested in narrow-scoped tests because they will provide important, constructive information about the learner's development of skills and knowledge. However, if the test is too narrow, the results are not meaningful for teaching or learning!³ A narrow-scoped measure of literacy would test the reading, vocabulary and computational skills needed for a specific job or activity. Instructors teaching functional literacy will probably want to use a narrow scope of measurement. According to Popham a narrow-based test can allow for specific, often trivial, focus in teaching. For example, learners could be tested for literacy by asking them to sign their names. In contrast, a broad-based test can be so broad it doesn't measure accurately any of the important behaviors to which the learner is directed. For example, learners who have been taught how to balance their check books might fail a test in business math. Conversely, learners who have been taught to do only basic math may not transfer these skills to a functional task.

On the other hand, even if providers and taxpayers are not opposed to narrow-scoped tests, they might prefer to see wide-ranged, broad scope tests. Decisions on this issue relate to the types of behaviors that are to be measured and the systematic treatment of the behaviors in the testing process. A broad based measure of literacy is a test that would measure the subjects' skills in general reading, vocabulary and computational skills needed for daily activities. Instructors who are teaching general literacy will probably want to use a broad based test. Clearly the scope of the material being taught should dictate the choice of testing scope.

Many suggest that program managers prefer to see broad-based tests while teachers prefer to see narrow-based tests. So some of the issues are related to the perspectives of the persons advocating one approach or another.⁴

APPROPRIATENESS TO TESTING GROUPS

Test questions, scoring scales and organization should be appropriate to the population tested. Most particularly, tests

³ W. James Popham, *Modern Educational Measurement*, Prentice-Hall, Englewood Cliffs, New Jersey, 1981, page 57.

⁴ *Ibid.*, pages 56-58.

should be neutral and unbiased to the range of ethnic, racial, age and gender groups that are expected to take the tests.⁵

TYPES OF TEST SCORES

Test measurements or scores should be pertinent and viable to the intended uses of tests. Just as distance can be measured with a variety of different scales (e.g. yards and meters), skills can be measured by a number of different scoring systems.

Test measurements or scores should be pertinent and viable to the intended uses of tests. Just as distance can be measured with a variety of different scales (e.g. yards and meters), skills can be measured by a number of different scoring systems.

For the most part, the suitability of test scores will be determined by user frames of reference. Put differently, scores and measurements will be most meaningful and usable if they reflect the "real world" concern of test users rather than abstract numbers. For example, test scores may reflect whether test takers can read, perform the duties of a job, or possess skills equal to most adults.

In most cases, the selection of one frame of reference should not preclude use of other frames of reference. Test scores can be "crosswalked" to determine equivalent scores to other frames of reference, and scores can be analyzed in a variety of ways (e.g. performance on a skill level scale can be viewed in absolute terms or in terms of how the test score compares to those of other test takers). However, the primary scoring system of a test commonly reflects a focal frame of reference, and this frame of reference becomes the primary perspective that guides use of the test.

Among the frame of reference issues of concern to users of adult skill assessment tests, three are immediately apparent:

- School Grade Level Reference
- Fixed Level Improvement Reference
- Criterion- or Norm-reference

These three issues will be discussed in the following pages.

School Grade or Other Scales

There is some disagreement concerning how adult skill tests should be scored. There are two basic choices: (1) use the interpreted academic grade level scores from tests, or (2) use scores that reflect another frame of reference for adult competencies.

⁵ *Standards for Educational and Psychological Testing*, American Psychological Associates, Washington D.C., 1986, pages 23-25.

- **Grade Level Scores.** Many grade level tests have conversion tables that are based on norms of school children. Grade level measurement is advantageous because it allows comparison of many different learners, not just special populations such as adult education. It's also easy for most people to understand. For example, that a learner is reading at the fourth grade level has general meaning to most people. However, some research suggests that adults may not perform as well on tests that measure skills in terms of academic grades. One limited study found that marginally literate adults may not be meaningfully measured on auditing and reading scales by grade level when compared to children. This study suggests that the performance of adults may be confounded by other factors in the testing process. Consequently, measurement by grade level may not be an accurate measure of performance for adults.⁶ Another major disadvantage of using grade level is the potential negative impact on the self-esteem of adult learners.
- **Other Scores Or Transformations.** Test performance can either be reported as a raw score or it can be converted to more meaningful indicators. One testing system, for example, uses scattered test scores to determine 'A', 'B' and 'C' achievement levels in a graded hierarchy of complex skills. The skills are based on what are thought to be adult level proficiencies rather than academic skills found in grade schools.⁷ Reporting percentile ranks from raw score data in program performances can be useful since it will provide comparative information that is based on the distribution of averaged program outcomes. More sophisticated procedures can also be developed that would allow a learner's performance on one test to be forecasted with the score on another test. Such a system of scoring would be helpful in that it would allow for interpretations across similarly scored and administered performance measures.

In many cases, it is not necessary to make a final choice between grade level and other scores.

In many cases, it is not necessary to make a final choice between grade level and other scores. The construction of crosswalks, sometimes referred to as correspondence scales, based on large samples of representative persons taking several tests, can establish a reliable correlation of scores between

⁶ Sticht, *Op. Cit.*, February 1982, ERIC ED217-168, pages 5-21.

⁷ Comprehensive Adult Student Assessment System, SDCCD Foundation, San Diego, *Administration Manual*, September 1989, pages 4-7.

tests.⁸ Since educators in different classes may want to use different measurement scales, it may be useful to develop a crosswalk that would allow one unit of measurement to be converted to another.

Fixed Scale or Improvement Scale

Regardless of the choice of scoring scales, a learner's progress can be measured against an expected level of performance or against an expected rate of improvement. Two types of performance scales are considered in this section. The first is a fixed scale in which the learner is evaluated according to an expected level of achievement. The second is a relative scale in which the student is measured by his or her own change in performance. These are summarized below:

...a learner's progress can be measured against an expected level of performance or against an expected rate of improvement.

- **Skills Relative to a Fixed Standard.** With this approach learners must achieve a specified minimum level of performance to be considered successful in their course of study. For example, brain surgeons and air traffic controllers should pass absolute standards. However, problems arise when such scales are used in adult education. Although there are many instruments that claim to test for absolute standards, it is very difficult to get agreement on the standards in fields such as literacy. Another problem with this approach is that it does not recognize the limitations that all humans bring to their tasks.
- **Skills Measured in Terms of Improvement.** This approach compares a person's performance at the beginning and at the end of the class. Standards attached to this scale of measurement set the rate of change as the goal. Clearly, this is not the way to test airplane pilots. However, it does offer some benefits for adult education. The change between the pre/post test scores can be compared with other programs or against rates of change standards. With a relative standard in which individuals see themselves changing the attention is on personal growth rather than an arbitrary absolute standard set by someone outside their social environment. This approach is more likely to promote the self-esteem of learners than the requirement that each meet a minimum standard to be successful. It is also a fairer measure of performance for programs dealing with learners who have significant and diverse skill deficiencies.

⁸ Tom Sticht has established rudimentary correspondence scales among several basic skill assessment tests (see Thomas G. Sticht, *Testing and Assessment in Adult Basic Education and English as a Second Language Programs*, Applied Behavioral and Cognitive Sciences Inc., San Diego, California, January 1990, pages 33-36).

Some tests focus on the measurement of a specific performance such as the calculation of square roots. The learner either performs or does not perform the operation. Other tests interpret performance in terms of how others perform on the same test.

The choice between a fixed or a relative scale largely depends on the objectives of the class. A fixed scale is necessary to certify minimum qualifications for a professional license such as plumbing; a relative scale is more appropriate for instruction focusing on personal improvement such as literacy tutorials or remedial programs. It is also important to note that they need not be mutually exclusive—the fixed or the relative scales can be produced from the same test results.

It is possible to measure using both the fixed and relative skill levels with the same test. As long as there are pre/post tests, fixed performance measures can be derived from either test and improvement scores by taking the differences between the pre/post measures.

Criterion- or Norm-Reference

The referencing of test scores is also important. Some tests focus on the measurement of a specific performance such as the calculation of square roots. The learner either performs or does not perform the operation. Other tests interpret performance in terms of how others perform on the same test. The former is a criterion-referenced test while the latter is a norm-referenced test:

- Criterion-referenced tests are useful for measuring one's acquisition of a hierarchy of skills and behaviors—one's mastery of accomplishment. This type of test indicates whether the student can carry out specific tasks correctly. The individual's score can be tabulated as the number correct against the total number available. Test takers are not measured against peers but against an ordered scale of accumulative correct scores. For this reason, criterion-referenced tests are viewed as especially suited to competency based education.
- Norm-referenced tests compare an individual's performance to those in a normed or reference group. When considering the adoption of a normed test, it is important to note who is in the reference group. If the normed group is made up of others who are in adult basic education programs, then their average skills are the implied goal of the program. If the normed group consists of a representative sample of the entire adult population, then the implied goal of the program is the skill levels of the average adult.

Criterion- and norm-referenced tests are two basic categories of testing used in educational measurement. It is important to identify what each of these tests can do for the test user and the learner. It is generally agreed that they both have

a role to play in the effective measurement and interpretation of learners' and programs' performances.

The use of either type of testing method has implications for how the test is to be used. When the decision maker must choose among a group of applicants for a limited number of opportunities, a norm-referenced test may help to identify those who are performing the best among those who took the test. On the other hand, there may be situations in that the individual must be certified for skills, and in that testing situation a criterion-referenced test is best. For guidance decisions, norm-referenced tests seem to have an advantage; but in program evaluation, it seems that criterion-referenced tests are best. Criterion-referenced tests seem to have the greater power for assisting the process.⁹

For the most part, decisions about the scores of tests are not a matter of how tests measure skills, but how test scores are reported and whether crosswalks exist that allow scores to be compared to other tests or frames of reference.

3.4

Documentation of Validity and Reliability

Validity and reliability are two of the most important criteria for evaluating tests.¹⁰ These two elements are interlinked:

- Validity is the extent to which a test measures what it claims to measure. For example, a test that claims to measure basic math will be invalid if it only measures capacity to conduct addition and division, or uses written questions that prevent respondents with literacy problems from demonstrating their math skills.¹¹
- Reliability is the ability of the test to measure what it purports to measure consistently among respondents or over time. For example, a literacy test would have poor reliability if test scores varied greatly among

⁹ W. James Popham, *Modern Educational Measurement*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1981, pages 35-38.

¹⁰ *Standards for Educational and Psychological Testing*, American Psychological Association, Washington D.C., May 1987, page 9.

¹¹ Some social scientists divide validity into internal and external validity (see Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research*, Rand McNally College Publishing Company, Chicago, Illinois, page 5-6). This section deals with "internal validity," "External validity," which is the extent to which measured results can be generalized or compared to larger groups, will be discussed in a later section dealing with the criteria of representative and comparable data.

respondents who have equal levels of literacy, or if a test that originally indicated high literacy levels revealed low literacy levels when taken by the same group three weeks later.

Both validity and reliability are determined not only by test design, but the procedures by which tests are administered. These two interconnected criteria are discussed below.

VALIDITY

All tests should provide rigorous documentation of their validity. A panel of experts convened by national associations concerned with the quality of tests and measures recommended that test users review documentation of three types of validity.¹² These include:

- Construct Validity
- Content-Related Validity
- Criterion-Related Validity

Validity and reliability are important criteria for evaluating tests. Validity is the extent to which a test measures what it claims to measure. Reliability is the test's ability to measure consistently.

In many ways, these three types of validity are interrelated and overlapping. The nature of each of these types of validity and approaches for documentation are discussed in the next three subsections.

Construct Validity

Construct validity addresses whether a test effectively measures a theory or configuration of traits about human conditions or behavior. For example, a test about political philosophy would have to define the nature of political philosophies being measured (e.g. liberal or conservative) then develop a battery of test questions that provide a "valid" overview of whether test respondents fit into the specified "constructs" of political philosophies. Another example is testing for a "concept" of literacy. Given that literacy is conceptualized into a theoretical construct (e.g. literacy norms of populations, job-related literacy, literacy skills allowing performance of specified tasks), the construct validity of a test would be determined by whether the test truly measures whether respondents have the skills to be literate based on the concept of literacy being tested.¹³

¹² *Standards for Educational and Psychological Testing*, American Psychological Association, *Op. Cit.*, pages 9-18.

¹³ While the nature of documentation for construct validity will vary from test to test, users of tests should look for a number of key indicators. Statistical correlation of test results with other tests measuring the same or related "theoretical constructs" is an important form of documentation. (It is critical that such correlations be conducted on test results from the same or similar populations of test takers.) Another valuable test entails the internal correlation of responses to questions in the tests that are assumed to measure the same or related conditions or behavior. In some cases, assessments of construct validity are made by a "jury of independent experts" who review and evaluate the theoretical logic, consistency and balance of tests. Test users should also beware of using tests for

Content-Related Validity

Content-related validity deals with whether the content and balance of the test measures the conditions or behaviors to be measured. Content-related validity overlaps with construct validity, but differs in that it focuses on the relationship of specific nature of questions to the specific types of conditions or behaviors to be measured. For example, content-related validity might address whether a test on basic computer operations reflects the actual skills and knowledge needed to use a computer. The concern here would be whether all the essential skills were addressed in a balanced fashion. Content validity would require that all "essential" skills be measured. It would also address whether the balance of questions reflects the balance of required skills (e.g. there would *not* be 30 questions on the keyboard and only one question on operating systems).¹⁴

Content-related validity might address whether a test on basic computer operations reflects the actual skills and knowledge needed to use a computer.

Criterion-Related Validity

Criterion-related validity addresses whether tests effectively measure how well respondents meet specified conditions. It is interlinked with the previously discussed types of validity, but focuses on whether the test has the capacity to measure the degree to which respondents meet a specified condition. This standard is also related to previously discussed criteria dealing with the range of measurement (Criteria 3). However, it deals with whether the measurements within the testing range are dependable, not the breadth of the range. For example, criterion-related validity would deal with whether 30 questions on a literacy test would accurately measure the test taker's reading proficiency on a scale that ranges from 6th to 12th grade levels.¹⁵

purposes that are different from the original purpose of the test (e.g. using a math test to assess computer skills). Finally, attention should be given to specified procedures for administering the test, and how well these procedures are implemented in actual testing situations. Procedures should not bias respondents toward certain responses, unduly prepare respondents for the test (e.g. teaching for the test questions rather than overall test goals), or create uneven conditions or distractions for test takers.

¹⁴ Documentation of content-related validity is commonly interlinked with tests of construct validity. Prospective test users should look for two types of documentation. First, there are a variety of approaches for observing the actual conditions or behaviors to be measured by tests, then assessing whether they match the content of the test. Second, it is important to review the context and format of test questions to ensure that they are relevant (e.g. terms or pictures about computer equipment should be current or similar to that actually being utilized by test takers).

¹⁵ Documentation of criterion-related validity can be provided in a number of usages. Scores from one or more tests can be correlated to assess whether tests measuring the same range of the same condition are comparable for similar groups of test takers. Parallel responses would support validity, while variations between two or more tests would suggest that at least one test has shortcomings. In some cases, followup data can be used to determine if test takers measured to be a given proficiency level ultimately receive remedial instruction at levels lower than their test scores. Finally, attention can be addressed to assessing the error level of decision theory associated with test scores. All measurements have some range of error; therefore, decisions made on measurements will have some errors (e.g. a certain percentage of students judged to be capable of 12th grade level instruction will not be capable). The error level of test-based decisions indicates criterion-related validity.

RELIABILITY

Documentation of test reliability is also critical. Reliability, the capacity of a test to measure consistently over time or among test takers, is determined by both test design and test administration.

Testing procedures, and the actual application of testing procedures, is a major determinant of reliability. Test users should make sure that tests are accompanied by testing procedures that ensure that respondents will take tests under equal conditions.

Test Design

Test questions and the overall structure of tests are major factors determining reliability. If questions are ambiguous or have different meanings to persons from different social groups, test results are not likely to be reliable. For example, if questions dealing with literacy skills place undue emphasis on farm animals, respondents from inner cities may be less likely to score as well as equally literate test takers from rural areas. Such a test would have low reliability for urban groups. The order and combination of questions can also impact reliability. This is particularly important in cases where there are pre/post tests to determine whether test takers have changed over time (e.g. improved skills, altered attitudes). Under such conditions, familiarity with the tests may be as much responsible for changes in scores as actual changes among respondents.¹⁶ In such cases, pre/post tests should have a rotating battery of different but equally valid questions. This ensures that test takers will not remember answers to questions from the pre-test.

Test Administration

Testing procedures, and the actual application of testing procedures, is a major determinant of reliability. Test users should make sure that tests are accompanied by testing procedures that ensure that respondents will take tests under equal conditions.¹⁷ For example, test takers at different times and locations should not be exposed to different levels of distractions such as noise or room climate. Similarly, all test takers should receive the same instructions and proctoring. Under most conditions, test takers should also be given the same amount of time to complete the test. Tests should include understandable instructions concerning test administration, and ideally procedures for documenting the application of these procedures in actual testing circumstances.

¹⁶ Donald Campbell and Julien Stanley, *Op Cit.*, pages 5-6.

¹⁷ Respondents themselves may not approach tests with equal personal orientations. For example, some test takers may be more anxious or fatigued than others. However, test procedures should be designed to ensure that the external environment under which respondents take tests is as equal as possible (*Standards for Educational and Psychological Testing, Op. Cit.*, page 19).

3.5

Representative and Comparative Data

Test results are not used in a vacuum. They are commonly interpreted in reference to comparison groups, and increasingly used to analyze the status of tested populations. Test users should therefore consider the following three characteristics.

- Capacity to Generalize Aggregate Results
- Capacity to Generate Norms
- Capacity to Measure Improvement

The above three capacities generally depend more on testing procedures than test design. Each of these three capacities is elaborated in the following pages.

CAPACITY TO GENERALIZE AGGREGATE RESULTS

Concern over program performance and the status of large populations adds an important new dimension to testing. They must address the issue of whether test results can be aggregated to generalize the participants of programs or specified populations.

The capacity to generalize test results to populations targeted for testing is becoming increasingly important. The need for this capacity stems largely from pressures to evaluate the performance of social programs of standardized measures of their impacts on participants, and to assess the conditions of our general population. This is particularly true for education and training programs. Specifically, there is a growing need to measure the results of education in terms of whether students reach* average societal competencies or show notable skill improvements. Similarly, there are needs to assess whether our total population or specific sub-populations have achieved the skills and knowledge that is required to function effectively in today's society.

Concern over program performance and the status of large populations adds an important new dimension to testing. In the past, testing and test results have focused on the individual. In the future, tests will continue to focus on the individual. However, they must also address the issue of whether test results can be aggregated to generalize the participants of programs or specified populations.

This growing need to generalize test results adds to the criterion of creating representative data. For the most part, this is a matter of test administration rather than test design. More specifically, it is more a matter of social science and survey

methods than the testing and measurement sciences that have tended to dominate the development of testing systems.

This section will provide an overview of the conditions necessary to collect representative test data. First it will provide an overview of the two major approaches for collecting representative test data, namely census and sampling. Second, it will elaborate the conditions required for representative samples.

TWO APPROACHES TO REPRESENTATIVE DATA

There are two approaches to collecting representative test data. The first is to test everyone in the population of interest. This is called a census. Taking a census is not always possible. As a result, the main challenge becomes getting a sample with representative data on test scores.

There are two general approaches to collecting representative test data. The first is to test everyone in the population of interest. This is called a census. If a subgroup of the population is tested, the result is a sample. In many cases, taking a census is not feasible. As a result, the main challenge becomes getting a sample with representative data on test scores. In this section the merits of a census and a sample will be discussed and the critical importance of representative data will be explained.

Census

Collecting census data can be a powerful method of data collection if the test takers are in a controlled system of education, such as the Armed Forces with enforced test taking. Civilian adult learners, however, can enter or leave the classroom whenever they choose. Census test data can be collected by requiring all students who enter programs to take tests at the start or soon after the start of their entry. While this is one way to ensure representative entry data, it will not solve the problem created by those who leave without taking the followup tests. Thus, if the program evaluators attempt to estimate the changes in performance of those who remain, they will be providing only information about those learners who went through the entire program.¹⁸

Significant errors can develop in census data due to inadequate resources for quality control. For example, educators who try to repeat a form of the entry test for every learner may not have enough time to make sure that every student takes the test after the same time interval and under the same administrative procedures. As a result, the comparability of test scores would not be truly reliable.

¹⁸ One possible solution is to invest in a high quality tracking and monitoring system for those who leave early.

Sample

An alternative design is one that relies on high quality probability sampling procedures. In most cases, this approach has significant advantages over the census because it is cheaper and more efficient. If appropriate procedures are not carefully followed, however, the data that is collected can be more misleading than helpful.

Probability sampling has significant advantages over the census because it is cheaper and more efficient. However, if appropriate procedures are not carefully followed, the data collected can be more misleading than helpful.

A representative sample requires a sampling design that ensures that collected test data is an accurate reflection of the total group of learners.¹⁹ High quality, representative data can be collected through probability sampling procedures. If these plans are carried out effectively, the resulting data costs less than a census, and random error is the major problem in the interpretation of the data. Since random error can be calculated, the data will be much more useful than the results obtained from a poor design that has haphazard errors that cannot be estimated.

REQUIREMENTS OF REPRESENTATIVE SAMPLING

The primary tasks of sampling are to collect data from a representative portion of a population that is large enough to produce acceptable margins of error and allow analysis of subgroups of the sample.

To emphasize the dangers of collecting data from a poorly constructed sample, consider the fate of the *Literary Digest*. During the 1936 Presidential campaign, the *Literary Digest* sampled millions of telephone subscribers and auto owners. It picked Republican Alf Landon to be the winner by 10 percent. Franklin Roosevelt won by a landslide.

Interviewing only auto owners and telephone subscribers during the depression was a fatal error. Why? They tended to be more "well-to-do" than most voters. The *Digest* failed to consider this bias in the sample they surveyed to represent the entire electorate. They also accepted an 85 percent refusal rate in a mail sample. The *Literary Digest* thought that all they needed was a lot of responses to get accurate and reliable results. In reality, only people who were wealthy enough to own a car and pay for return postage responded.

¹⁹ Researchers generally divide data collection plans into two types: (1) formal and planned data collection and (2) informal and haphazard. If data is collected according to plans that are designed to be representative of the underlying groups, then formal and planned data collection methods have been put in place and implemented. There are, unfortunately, numerous examples of data collection activities which yield unrepresentative or biased information as a result of informal and haphazard data collection methods. For example, educators might allow learners to take as much time as they need to complete their basic skill tests. The problem in evaluating such data is that there is uneven, haphazard administration of the tests and the data that is produced is influenced by the different testing procedures. Thus, some students who have to go to work don't score as well because they have to leave earlier than the others. It does not necessarily follow, however, that these students are less skilled. It is important that common procedures be used for administration and proctoring of tests in the various sites.

During the same 1936 Presidential campaign a relatively unknown polling group focused their resources on getting a high quality set of data. Although they used a much smaller sample, 1,500 respondents in a quota sample, their forecast was correct. It was the beginning of a success story for the Gallup Poll and representative sampling techniques.

Representative Samples Are Critical

Representative data hinges upon random sampling. This approach uses blind selection procedures where everyone in the target group has an equal chance of being selected.

The most important decision when collecting data to represent a larger group, is *whom to sample*. It is important that the sample selected be representative of the total population (target group) of interest.

Representative data hinges upon *random sampling*. Random sampling uses blind selection procedures where everyone in the target group has an equal chance of being selected for an interview. Therefore, every person in the target population to be studied will have a chance to be in the sample.

To illustrate the dangers of "non-random sampling," assume that a phone survey is being conducted, and that the target group is all students who graduated from a specific adult school. If data are collected from a "convenience sample" that saves phone toll charges by interviewing only those who are in the local call area, the sample will *not* be random. Maybe those who moved out of the area got better jobs because they were willing to move. The results of the interviews could lead you to conclude erroneously that their education did not help them get jobs. However, if each completer was assigned a number and a few of the numbers are blindly drawn "out of a hat" (or a computer), a random sample will be created to provide representative data.

It is also important to recognize that representativeness is far more important than the size of a sample. To avoid the *Literacy Digest's* fatal error, tests should be administered to a small sample that is randomly selected to be representative rather than a large number that is not representative. This will yield better information than thousands of poorly collected responses because:

- ~ More care can be given to ensuring that the sample of respondents is representative.
- ~ The process is generally less expensive.
- ~ It is possible to make certain that all persons to be interviewed are interviewed.
- ~ Care can be taken to ensure accurate responses.

Sampling is not always appropriate, but getting representative data is!

Sampling Error

Truth is an illusive goal. Scientists, whether they study physics or society, have long known that when we seek to measure reality, we never really get an absolute figure. Whether we measure atomic particles or unemployment rates, there is always some give and take in our numbers.

The amount of "shake" in our numbers is reduced as the number of observations increases. Imagine that you are sampling all the coin flippings in the universe. For this infinite number, you know that the average will be 50/50 for heads and tails. But this average will vary with the number of flips you make. If you take a "sample" of ten flips, the results could easily be seven heads and three tails. But if your coin flipping sample has 100 tosses, the chances are much greater that the results will be closer to the 50/50 distribution for the universe.

Thus, test users should be keenly aware that the amount that scores collected from a randomly selected sample differs from the population of interest will decline as the sample size increases. Data from survey samples will always have some error. Unless you take a census, which "samples" everyone in your target population, and collect data from everyone in the sample, survey data will vary somewhat from the actual population. This is called *sampling error*.²⁰

While it's always good to have large samples, there are points where the costs of added data collection do not result in significant reductions of sampling error. Thus, it is important to recognize that after a certain point, the reduction of error

²⁰ Fortunately, we have tools to measure sampling error. Statistical techniques can tell us the range of error that our survey data is likely to have and the probability that this range of error is true. In statistical jargon, the range of error is known as *precision* and the probability that this range of error will not be exceeded is known as *confidence*. Examples can illustrate these concepts.

Suppose there are 5,000 employers in your county and you randomly select 500 to be interviewed. Results indicate that the average employer hired 10 new workers last year. How close is this estimate to the average income of the entire group of 5,000? Precision refers to how close the estimate is to the entire survey group's average income. If the estimate is made with an error margin of plus or minus 5 percent, it is more precise than if the estimate is made with plus or minus 10 percent error margin. Greater precision requires either reductions in confidence or increases in sample size.

Confidence indicates the likelihood that your estimate of precision is correct. Professional survey research commonly uses a 95 percent confidence level as the standard. This means that 95 out of 100 times the sample you are working with is representative of the population of interest. In other words, the chances are 95 out of 100 that survey responses mirror the population you are studying within a given range of precision.

It is important to note that you can make trade offs between precision and confidence levels. Given a sample size, higher standards of confidence will reduce the probability that your data exists within the range of precision you have specified. Exhibit 2.1 presents the precision estimate for seven sample sizes and two levels of confidence. Notice how the precision increases as the confidence level decreases. Notice also how the precision ranges get smaller as the sample sizes increase within each of the confidence levels. Exhibit 2.1 also illustrates how sample size affects precision and confidence.

Say that there are 50,000 Californians who have completed a high school class in retail marketing. You take a random sample of 1,000. You manage to get a response out of every one of them, and 50 percent agree that the course was no good because the methods they were trained in were not effective. According to Exhibit 2.1, the chances are 95 out of 100 that between 46.9 and 53.1 percent of all 50,000 of the completers would agree that the course was no good because the methods were ineffective. However, if we wanted 99 percent confidence, we would have to say that between 45.9 and 54.1 percent agreed that the course was no good.

gained by increasing the size of random samples is not worth the added costs of data collection.

For example, with a sample of 250, your results can be expected to vary somewhere between plus or minus 6.2 percentage points from the total population. For a 1,000 respondent sample, this range of error is reduced to plus or minus 3.1 percentage points, and for a sample of 5,000 it is plus or minus 1.4 percentage points (Exhibit 3.3). For those of you with an eye to confidence levels, these figures were calculated for 95 percent confidence.

Exhibit 3.3

**SAMPLING ERRORS WITH RANDOM SELECTION
AT TWO CONFIDENCE LEVELS
AND SEVEN SAMPLE SIZES**

Sample Plus Size:	Percent Plus or Minus at 95% Confidence:	Percent Plus or Minus at 99% Confidence:
10,000	1.0%	.3%
5,000	1.4%	1.8%
2,500	1.6%	2.6%
1,000	3.1%	4.1%
500	4.4%	5.8%
250	6.2%	8.2%
100	9.8%	12.9%

*While it's
always good to
have large
samples, the costs
of added data
collection do not
result in
significant
reductions of
sampling error.*

Most survey professionals set their sample sizes between 500 and 2,000. The upper numbers of this range are generally chosen to allow analysis of groups within the total sample (*i.e.*, views of women and minority respondents) rather than to decrease the sampling error for the total sample. The wonderful thing about probability sampling is that these guidelines are generally true regardless of the size of your population of interest.

Importance of Response Rates

High response rates are a critically important principle of sampling. If persons are selected to be part of a representative sample, it is extremely important that test scores be collected from them.²¹

²¹ The response rate is found by dividing the number of completed interviews by the number of expected interviews. For example, if you drew up a list of 100 people to be interviewed and only managed to interview 60, your response rate would be 60 percent.

Higher response rates are a critically important principle of sampling. If persons are selected to be part of a representative sample, it is extremely important that test scores be collected from them.

The primary goal of probability sampling is to collect a sample which is representative of the population of interest. As noted above, the key rules for such sampling are that every member of the population of interest must have an equal chance of being selected for the sample. It is equally important that sampled persons in the least do not select themselves out of the sample (i.e. not responding or refusing to respond). Thus, every effort should be made to collect data from all selected persons in the sample to avoid self-selection bias. In this manner, the sample is thought to reflect only errors due to chance and not those due to the selection judgments of data collectors or the self-selection of respondents.

There are many causes of low response rates. In some cases, a low response rate is due to lazy data collection. It may not be possible to locate a potential respondent because he/she has left town with no forwarding address. A low response rate may also be due to the failure of respondents to answer all of the questions. The "hard-core" refuser also contributes to a low response rate. Some of the sample will not want to take time from their busy schedules to be tested.

Why worry about low response rates? Because the views of non-respondents may be dramatically different from those of respondents.

To understand the detrimental impact of a low response rate, consider the maximum distortion that might result (Exhibit 3.4). Say you took a sample of 1,000 graduates and you had a non-response rate of 40 percent. Fifty percent of the sample said that their high school education had improved the quality of their lives, and 50 percent said that it was not beneficial. Exhibit 3.4 indicates that the maximum range or possible error for a response rate of 60 percent is 44.6 percentage points. With a 60 percent response rate, it is possible that the views of the entire population from which you took your sample is that between 27.7 percent and 72.3 percent feel that high school improves the quality of their lives. This range of possible viewpoints provides little solid information for interpretation.²²

The importance of getting high response rates cannot be overstated. For the most part, survey professionals commonly assume that the views of non-respondents parallel those of respondents. However, the viability of this assumption disintegrates rapidly as non-responses climb.

²² Half of 44.6 percent is 22.3 percentage points. Add or subtract 22.3 percent to the 50 percent who said that their high school education did improve the quality of their lives. If the views of non-respondents are the opposite of respondents, the real views of the population could vary from an upper value of 72.3 percent who state that they found high school beneficial and a lower value of 27.7 percent.

Exhibit 3.4
INCREASES IN ERROR MARGIN WIDTHS AT
DIFFERING NONRESPONSE RATES

Assumptions: Number = 1,000
50% respond "yes"
50% respond "no"

Response Rates (Percent of Total Sample)	Total Standard Widths of Error (Percent Variation)
100	3%
90	16.0%
80	25.6%
70	35.1%
60	44.6%
50	54.1%

SOURCE: This table is adapted from William G. Cochran, *Sampling Techniques*, 3rd Edition (John Wiley & Sons, New York), 1977, page 364. See also Carol H. Fuller, *Weighting to Adjust for Survey Nonresponse*, *Public Opinion Quarterly*, Volume 38, 1974, pages 239-246.

CAPACITY TO GENERATE NORMS THROUGH COMPARISON GROUPS

Comparative data is important because it establishes norms for evaluating skill assessment scores.

Comparative data is important because it establishes norms for evaluating skill assessment scores. For learners and counselors, comparison scores (e.g. average for general population, other test takers, etc.) provide a reference point for interpreting individual scores. For those using scores to evaluate program performance, comparative data provides a basis for assessing the success or failure of programs relative to other programs or modes of instruction, and how well learners with varied backgrounds perform within different types of programs. For those seeking to use test scores to certify skill attainment, comparative data provides a "reality" check on the extent of instruction likely to be required to bring the average learner to desired levels of proficiency. Thus researchers, program managers, teachers and learners benefit from good comparative data on performance.

Test users should be concerned with five major types of comparisons. These are: (1) learner and other test takers, (2) learner and reference groups, (3) beginners and completers, (4) completers and non-completers, and (5) learners with different characteristics.

Comparison to Other Test Takers

One important comparison is of individual test scores to the scores of all other test takers. There are many technicalities with this comparison, most of which will be discussed in subsequent pages dealing with other types of comparisons. However, four are particularly noteworthy to prospective test users. First, comparisons to all test takers requires that procedures must be provided for producing aggregate data on all test takers (or a sample of all test takers). Second, the characteristics and timing of past test takers should be defined (e.g. number taking tests in past years, characteristics of test takers in past years, etc.). Third, procedures for providing data on all test takers should ensure that summary data is truly representative of all test takers. Fourth, attention must be given to whether the same or comparable tests are administered in the same fashion to learners. The administration of different and non-comparable tests is likely to lessen the value of summary data for each test because it will not necessarily reflect any given group or program.²³

Comparison to Reference Groups

Under almost all circumstances, skill assessment tests should be normed to allow comparison of scores to one or more reference groups. For the most part, these norms are likely to be established from two groups: test scores from a representative group of all test takers, or test scores from a representative sample of all adults. These two approaches are specified below:

- **Sample of Test Takers.** Tests can be normed on a sample of test takers. Test documentation should clearly specify the group of test takers used to norm the test. The reason for using tests is to evaluate programs and to guide and place learners. Irrelevant reference groups will not aid any of those processes. If the norms of a test are unrelated to the group being tested, or if the norms are not relevant to the purposes of teaching, it does not make sense to use the test. The characteristics of the group of test takers used to establish norms must reflect the characteristics of the population targeted for use of the test. If the normed group is a group of elementary school students, it does not make sense to focus the testing on adults using that instrument.

²³ Robert L. Ebel, *Essentials of Educational Measurement*, Third Edition, (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1979), pages 190-200.

- **Representative Sample of Adult Population.** Tests can also be normed on a representative sample of the adult population. Test documentation should carefully explain the selection of a representative sample of test takers, the procedures of administering the test to the sample, and the dates of test administration.²⁴

In summary, skill assessment tests should be normed to one or more reference groups, reference groups used for establishing norms should be appropriate to the intended uses of the test, and the procedures for establishing norms from reference groups should be clearly documented.

Capacity to Measure Skill Improvement

Skill assessment tests should be normed to allow comparison of scores to one or more reference groups.

Measurement of skill improvement among learners is one of the most important roles of skill assessment testing. This requires the administration of pre/post tests to allow comparison of learner skill levels at the beginning of program participation, and at one or more points after learners start programs. It also requires an inventory of interchangeable test questions which can be used to create new tests about the same skills so that post-test scores reveal skill improvement rather than test similarity. These two requirements are elaborated below:

- **Procedure for the Pre/Post Test Administration.** Test users concerned with measuring skill improvement should make sure that procedures exist for the administration of pre/post tests, or that the test can be used for pre/post tests. A critical issue in the administration of pre/post tests is whether to collect data from a sample or all students. If pre/post test data is to be used for individual assessment and counseling, it will be necessary to collect data from all students. If the data is to be used only for program planning and evaluation, a sample will be less expensive and more appropriate. In either case, there are serious obstacles to collecting post-test data from adults (e.g. many adults do not complete programs or attend regularly, adult instruction is frequently decentralized and irregularly scheduled). To maximize value to individuals and ensure representative data for program evaluation, users should ensure that procedures for the collection of pre/post test data are both feasible and reliable.

²⁴ Normed standardized tests are expensive, time consuming and subject to significant error in procedures. Consequently, users of tests must be concerned with the use of the normed conversion tables from raw scores or serious errors will result from their use.

- **Inventory of Interchangeable Test Questions.** It is very important that test scores on post-tests are not artificially high because respondents are familiar with the questions they answered in the pre-test.²⁵ The standard solution to this problem is to develop one inventory of interchangeable questions that measure the same skills. Development of such an inventory is no simple task. Test users should be sure that the interchangeability of questions is both valid and reliable, that procedures are carefully specified concerning the selection and ordering of questions to form new tests, and the range of error among alternative test batteries is clearly documented.

Measurement of skill improvement among learners is one of the most important roles of skill assessment testing. This requires the administration of pre/post tests

For specific information concerning the two issues above, test users should refer to previous sections of this report dealing with ranges of measurement, validity and reliability, and representative sampling.

Comparison to Those Who Leave Early

Students who leave programs may be different in important ways from those who stay. Thus, it may be important to assess the characteristics of early leavers and their skill attainment. Followup interviews may reveal why people leave a learning situation before completion. This information may be useful to educators who want to improve the completion rates in their classes. Test measures from those who leave early can also be compared to those who stay in the program to gain further information about programs.

Collection data from program learners presents many problems. However, followup or tracking systems can be developed. Research of this type is very costly to manage. Often adult learners who leave programs are difficult to find. They may move frequently. Some may change their names when they marry. In some households, the telephone service may be disconnected. Repeated calls are often necessary for those with a known phone number who are employed. Because of such problems, and the associated cost of this type of data collection, it is desirable to collect data from a sample rather than the total population of learners. Thus, a followup or tracking system to contact and collect data from a sample of learners must be well designed and carefully managed to provide cost-effective, high quality information about those who leave early.

²⁵ Donald Campbell and Julia Stanley, *Op. Cit.*, pages 5-6.

Comparison of Learners with Different Characteristics

It is highly desirable to collect test data that allows flexibility comparing test scores by the characteristics of individuals and the programs they use. This analysis requires test data to be individual rather than batched program data.

It is highly desirable to collect test data that allows flexibility comparing test scores by the characteristics of individuals and the programs they use. For example, it may be useful to see how the scores of minorities compare to those of non-minorities within different program settings. Or it may be desirable to isolate the individual characteristics of adult learners that are associated with success or failure to improve skill scores. These and other similar types of analysis require that test score data be coded and scored as individual rather than batched program data.²⁶ Test users should consider the importance of the flexibility to analyze data when selecting tests, and seek tests and accompanying data storage and retrieval systems which allow recall of test scores by individual rather than group averages within programs.

The five types of test score comparisons above are all desirable. However, they are not all essential. Capacity for conducting these and other types of comparisons should be prioritized with the intended uses of skill assessment tests.

The prior pages have isolated a number of pre-conditions for conducting the test score comparisons. These include: (1) storage of data for individuals, (2) use of a common test or tests with valid crosswalks, (3) capacity and procedures for pre/post tests, (4) collection of data that is representative of programs, (5) collection of data that is representative of the adult population, (6) followup or tracking systems, and (7) inventories of interchangeable test questions. There is an overlap between these pre-conditions and the capacity to conduct the comparisons outlined in prior pages (Exhibit 3.5).

Some of the pre-conditions for test score comparisons are essential for all types of foreseeable comparisons. Others are necessary for selected types of comparisons. Test users should make every effort to ensure that tests meet immediate needs, and that they have the capacity future use.

²⁶ Currently, data for many tests is aggregated into batch averages or ranges for classes or schools. Once this is done, it is impossible to recall data or conduct analysis based on individual characteristics. Capacity for storage and retrieval of individual test data is also important to the services proposed for the Strategic Recommendations for the EduCard_{SM} and Community Adult Education Information Services.

Exhibit 3.5
PRE-CONDITIONS FOR COMPARISONS
OF TEST SCORE DATA

Pre-Conditions:	Options:				
	Learner to Other Test Takers:	Learner to Reference Group:	Beginners and Completers:	Completers and Non- Completers:	Learners with Different Charac- teristics:
1. Storage of Data for Individuals	☒		☒		☒
2. Use of Common Test or Tests with Crosswalks	☒	☒	☒	☒	☒
3. Capacity and Procedures for Pre/Post Tests			☒	☒	☒
4. Collection of Test Data that Is Representative of Programs		☒			☒
5. Collection of Test Data that Is Representative of Adult Education		☒			☒
6. Followup or Tracking System			☒	☒	☒
7. Inventories of Interchangeable Test Questions			☒	☒	☒

☒ Required ☐ Not Required

3.6

Common Measurement of Skill Levels

It is extremely important that assessment tests provide a common "yardstick" for measuring skills and knowledge. Clearly, commonality *between* tests measuring literacy and parenting skill is not important. But commonality *among* tests measuring the same skills and knowledge is. From the standpoint of assessment and counseling, it is important that tests used by one program or locality be usable in another, and that tests administered at one point in time be comparable to

those administered at a later point of time. From the standpoint of placement and certification, it is important that those who use tests to certify skills and make placement referrals are making decisions based on common measurements that are also understood and used by those relying on skill assessment tests as a criteria for employment or acceptance into instructional programs. From the standpoint of program planning and evaluation, it is important to have common measurements to assess learning outcomes within and among programs.

USE ONE TEST

From the standpoint of program planning and evaluation, it is important to have common measurements to assess learning outcomes within and among programs.

The most obvious approach to ensuring reliable common measurements is to use one test for skill assessment. This is not to say that other tests cannot be used, but that one test be required to provide "core" or a common linkage for the multiple applications of skill assessment tests.

In many ways, use of one test is the optimal solution for developing a common denominator for assessment and counseling. In terms of the validity and reliability of test scores, use of one test greatly reduces potential errors in measurement. In terms of test use, one test reduces the danger that tests will be applied differently and promises lower operating costs through standardized testing procedures.

Unfortunately, use of one test may not always be desirable or possible. Educators and training programs may have long standing histories and financial investments in a variety of different testing systems. Many test users may need testing ranges or measurement detail that cannot be provided by one test.²⁷ Finally, there are likely to be a need for tests that utilize different modes of administration. For example, a "paper and pencil" test may not be applicable for a pre-literate or disabled learner.

TRANSLATE SCORES AMONG DIFFERENT TESTS

If one test cannot be used, the remaining option is to develop "crosswalks" that enable test users to translate the scores of different tests into common measurements. This option is generally viable and appropriate to a world with a significant plurality of testing systems. However, like the option of using one test, crosswalking test scores has limitations and shortcomings.

²⁷ It is, however, possible that one test might have options for more detailed tests and tests with overlapping ranges.

Test users should select or develop testing systems that meet all needs. If this is not possible, users should seek crosswalks with documented validity and reliability that allow translation of test results with a common scoring system.

There are three principle constraints to translating test scores. First and foremost, tests addressing the same general skill areas may, in reality, measure different things (e.g. tests measuring functional versus academic literacy). In such cases, translations of scores will be, at best, approximations and, at worse, totally inappropriate. Second, all translations (even those among similar tests) will add a new level of measurement error to the common scores. Different tests will be what they are—different. Crosswalks to produce translations will also be approximations—increasing the range of error in the translated test score. Third, use of score translations will add another level of administration and analysis to the use of test scores. This added workload can be reduced by computerized translations and analysis, but it will still add a new layer of complexity to the use of assessment tests.

The reality is that skill assessment, whether with one or many tests, will never be perfectly accurate or without administration problems. Where possible, test users should select or develop one test that meets all needs. If this is not possible, test users should seek crosswalks with documented validity and reliability that allow the translation of test results with a common scoring system.

3.7

Capacity for Analysis and Reporting

Capacity for test score reporting and analysis is an important consideration. Unfortunately, the design of many tests precludes the possibility of important types of analysis. In other cases, data storage and report production procedures do not allow or produce many needed reports.

Test users should consider three major issues dealing with the capacity for analysis and reporting:

- Nature and Relevance of Standard Reports
- Capacity to Analyze Individuals and Subgroups
- Access to Learners and Programs

The following pages provide a general overview of the above three issues.

STANDARD REPORTS

It is critical that the standard reports produced by tests be relevant to user needs and that they provide technical documentation about the limitations of interpretation. Most

Standard reports should reflect data and analysis needed on a routine basis

testing systems have procedures for producing standard reports. These may be produced by the testing system itself or take the form of procedures for test users to generate their own reports. Most skill assessment tests provide reports on individual test takers, and many provide reports on the test scores of groups and subgroups (e.g. all test takers, classes, racial ethnic groups by class).

Standard reports should reflect data and analysis that is needed routinely. It should not be necessary to undertake special analysis to obtain data that is required by users on an ongoing basis. Periodic standard reports should provide information on all major applications of test data (e.g. assessment and counseling, placement and certification, and program planning and evaluation). These standard reports should document the quality of information they present. Standard reports should include technical documentation about the liability and reliability of analysis, background on testing procedures, ranges of error, and other limitations on the interpretations of test data.²⁸

ANALYSIS OF SUBGROUPS AND INDIVIDUALS

Test users should also consider the capacity to analyze test scores for individuals and specified subgroups. This capacity should not be taken for granted. Many tests only store and report scores for batched groups of test takers (e.g. classes or schools), and do not report individual scores. Conversely many testing systems have limited capacity to generate scores for specified groups, a shortcoming that greatly undermines their value for setting norms or program planning. To maximize the uses of tests, users should select or develop tests that store data on individuals, collect demographic data on each test taker, and link scores from tests taken by the same individual over time.

- **Store Data on Individuals.** It is highly desirable that testing systems code and store test data on individuals rather than batched data on groups. This enables the recall of data to meet needs for individual assessment, counseling, placement and certification. It also allows data analysis by subgroups.
- **Collection of Background and Demographic Data on Individuals.** Where possible, test data on individuals should include basic demographic and background data (e.g. age, sex, ethnicity, classes, school, city). These data allow analysis of how

²⁸ *Standards for Educational and Psychological Testing*, Op. Cit., pages 83-84

It is highly desirable that testing systems code and store test data on individuals rather than batched data on groups.

subgroups perform and provide teachers and schools with feedback. If such descriptive data is not collected, the capacity to analyze the impacts of programs will be lost.

- **Linkage of Individual Scores Over Time.** Scores from tests taken by the same individuals over time should be linked. This allows analysis of how the skill levels of individuals and groups of individuals change over time. Some pre-conditions are required for such linkage. First, individual test takers must be assigned unique identification number when they take tests, and scores from subsequent testing be recorded by this same identification number. Second, a statewide database must be developed that allows test scores recorded at different locations to be accessed and analyzed.²⁹

The capacity to recall and analyze test scores for individuals and subgroups is fundamentally important to many applications of test data. The value of this capacity must be weighed against its costs.

LEARNER AND PROGRAM ACCESS

Access to test score data is also a major consideration for learners and programs. To the extent that it is economically possible and appropriate in terms of maintaining individual privacy, test data should be made readily available to users.

From the standpoint of the learner, individuals should be regularly informed about their test scores or provided with access to them for feedback on their educational progress. Such access should also allow individuals to provide documentation of their skill levels to educational programs and employers. Of course, individuals should only have access to their own scores and unidentified average scores to provide a context of judging relative achievement.

From the standpoint of authorized program staff, easy access to data is critical. Authorized counselors and teachers should be able to review the records of individual students. Program planners and administrators should be able to access aggregate data on groups to evaluate programs. Additionally, it is highly desirable that the testing system be capable of providing copies of test score databases with student

²⁹ The procedures and resources necessary to initiate unique learner identification numbers and a state database, are being explored within the context of option papers for other recommendations of the 1988-89 Strategic Plan for Adult Education (Dennis Porter, *Adult Education Access Card (EduCard₉₉)*; Policy Option Paper on Strategic Recommendation 4, and Dennis Porter, *Integrated Adult Education Database: Policy Option Paper on Strategic Recommendation 8*, Adult Education Institute for Research and Planning, Sacramento, California, March 1990).

identification deleted for purposes of special research and analysis.

All in all, test users should consider the nature of data and reports to be produced by testing systems, and the ease and flexibility with which users can access and analyze test data for various purposes.

3.8

Protection of Individual Rights and Privacy

Skill assessment and testing systems should incorporate safeguards to guarantee maximum individual rights and privacy for test takers.³⁰ Key considerations that should be addressed include:

- Explanations to Test Takers
- Options for Review and Retesting
- Safeguards for Privacy

These three considerations will be capsulated below.

EXPLANATIONS TO TEST TAKERS

Skill assessment and testing systems should incorporate safeguards to guarantee maximum individual rights and privacy for test takers.

Test takers should be fully informed of what tests measure, how test scores may impact them, and how the results will be used and reviewed. Understandable summaries and technical descriptions of the uses and limitations of assessment tests should be readily available to all test takers.

OPTIONS FOR REVIEW AND RETESTING

Test takers should have the right to fully review their test scores and, where appropriate, request retesting. In many cases, test scores will influence placement into instructional programs and provide certification for employment. Because of the potential impact of tests on individuals, procedures should guarantee test takers the right to expeditiously inspect their test scores, request retesting, request withholding of test scores due to alleged testing misconduct, and possibly expunge selected testing records.

³⁰ Ibid., pages 85-87.

SAFEGUARDS FOR PRIVACY

The procedures for testing and storage of test records should protect the privacy of individual test takers. Access to individual test scores should be rigorously restricted to individual test takers, authorized program staff, and persons explicitly authorized by the test taker to have access. Procedures and restrictions on the access of individual data should be clearly and explicitly stated. Penalties should be specified for the abuse of privacy rights. Files and databases developed for the storage of individual test data should be constructed with a variety of safeguards to prevent access by unauthorized persons. Finally, public reports on groups of test takers should not reveal information about individuals. Specifically, individuals should not be listed or identified with test scores, and breakdowns of test scores by individual characteristics should not allow identification of individual scores (e.g. a breakdown of scores by disability would reveal the score of a disabled test taker if there is only one disabled person in the group).

3.9

Cost Effectiveness and Administrative Viability

It is one thing to specify goals and standards for testing. It is another to implement a widespread skill assessment and testing system that serves intended purposes and meets acceptable technical standards.

To have meaningful impact, tests must be usable in diverse "real life" circumstances. It is one thing to specify goals and standards for testing. It is another to implement a widespread skill assessment and testing system that serves intended purposes and meets acceptable technical standards. Two issues are particularly important:

- Cost Effectiveness
- Administrative Efficiency and Viability

These two critical determinants of assessment and testing capacity will be discussed below.

COST EFFECTIVENESS

If tests are to be used with the technical rigor needed to provide information worth using, they must be worth the cost. It is desirable for the costs of testing to be low. However, the real issue is "cost effectiveness." Test users must ask whether the costs of using a test are greater or lesser than the costs of not using it. It may cost \$100 to administer and interpret a test for each learner—a cost that many might consider excessive. However, if the costs of instruction for the average learner is \$1,500, and a test can prevent faulty placement decisions for

one-third of all new students, this testing cost will be well worth the expenditure.

Test users should also consider the costs of inappropriate tests and low testing standards. It is important to keep in mind that test users will get what they pay for. If shortcuts are taken in administering tests, test scores may be invalid and actually worse than no testing at all.

Cost effectiveness requires more than merely selecting the cheapest test. Tests should first be reviewed in terms of a common criteria. Then the total cost of all acceptable tests should be compared. The total cost is measured by summing the following items: (1) the purchase price times the expected number of takers; (2) the cost of training test administrators, (3) the cost of proctoring the test, (4) the cost of scoring and analyzing the data that is produced from the test, (5) the cost of any consultants, such as a psychometrician, to conduct analysis of the test and its data, and (6) the cost of storing and maintaining the data bases that are produced.

Note that cost items two through four above are very important considerations. If teachers and learners are required to spend too much time collecting the data, they will probably find it difficult to perform their duties well. Thus it is possible that too much testing can be a disservice to the system as a whole.

ADMINISTRATIVE EFFICIENCY AND VIABILITY

It is possible that regardless of financial commitments, testing may be logistically unfeasible under some circumstances.

Given acceptable costs, or a willingness to pay high costs to get acceptable test performance, test users must also consider the administrative viability of tests. It is possible that regardless of financial commitments, testing may be logistically unfeasible under some circumstances. Under some conditions, testing may be disruptive, require excessive time, or undercut learner morale. The administration of tests may be logistically infeasible under conditions of decentralized instruction or diverse course schedules. Test users should carefully consider these and other factors when selecting or developing tests.

While barriers to testing should be fully recognized, it is also important to recognize potential solutions. Applications of interactive testing technology, self-administered tests, streamlined sampling techniques, and other devices can overcome many barriers. Barriers should be recognized. However, the existence of a barrier does not unto itself make testing infeasible.

3.10

Technical Auditing and Test Development

Tests should be evaluated not only for what they are but what they are becoming. Test users should select or develop tests that have institutionalized an open process of review and feedback, and that have long-term research and development plans. Three issues merit particular attention:

- Open Data for Review
- Third Party Evaluations
- Ongoing Research and Evaluation

These three issues are discussed in the following subsections.

OPEN DATA FOR REVIEW

To the extent possible within the constraints of protecting the personal privacy of test takers, testing data should be made available for review and evaluation. Potential users should have the option of assessing the quality of test data to ensure that the test performs in accord with test descriptions.

It is also highly desirable that test results be made available to the learner and authorized program staff through an interactive computer system.³¹

THIRD PARTY EVALUATIONS

Test users should seek and place importance on third party evaluations of the test. Important conditions to look for in such evaluations include:

- **Qualifications of Evaluators.** Evaluators should have advanced professional training and experience. Important areas of expertise should include testing and measurement, statistics, substantive background in to the subject areas of the test, and data processing.
- **Independence of Evaluators.** Evaluators should not be organizationally linked to the test, or have financial or professional interests in the test. Where possible, the evaluators should be retained by parties other than

³¹ The Community Adult Education Information Services and *Educard*TM proposed in California's 1988-89 *Strategic Plan for Adult Education* will ultimately require that test scores be stored in an Integrated Cross-Agency Data System.

the test providers. Provisions should be made to ensure that the evaluator is not subject to undue influence or pressure from the test providers.

- **Recent and Regular Evaluation.** Evaluations should be performed regularly to ensure that the current status of this test is always available. There should also be continuity of issues among past and current evaluations.
- **Balanced and Empirical Evaluations.** Evaluations should address all key issues in a balanced fashion and use professionally recognized empirical methods to assess test performance.

Evaluations will vary in content and approach from test to test. However, it is important that tests have objective evaluations that address a criteria of performance that is relevant to test users.

ONGOING RESEARCH AND DEVELOPMENT

Evaluations will vary in content and approach from test to test. However, it is important that tests have objective evaluations that address criteria of performance relevant to test users.

Skill assessment and testing systems are more than a set of questions and procedures. They are ongoing efforts to measure skills in a changing world. Test users must select or develop tests that are not only good today, but that evidence ongoing professional activity to make improvements to meet the needs of tomorrow. Attention should be given to evidence of past and current research and development, formal recognition of shortcomings matched by efforts to make improvements, prioritized agendas for future research and development, and a record of responsiveness and cooperation with clients.

In summary, test users should seek tests that are backed by staff and organizations that are open to questions, willing to undergo review and evaluation, and encourage an ongoing process of feedback and test improvement.

The ten criteria described above propose a descriptive checklist of elements that users might seek in skill assessment tests. Key themes from these interlinked criteria are that tests should be relevant to multiple users, that the quality of test the quality of test data must be acceptable or tests may cause more harm than good, and that tests should represent a systematic combination of procedures, applications and services that meet the needs of users. The next chapter formulates key elements of these ten criteria into recommendations that can be modified and prioritized for use in selecting or developing tests.

Chapter 4

Proposals

This chapter discusses fourteen proposals that will lead to the development of quality standards and performance measures for adult education.

This chapter presents proposals for the development of quality standards and performance measures for adult education. These proposals have been divided into two groups:

- (1) Priorities and Uses
- (2) Criteria for Skill Testing Systems

The first two of the four proposals in the Priorities and Uses section focus on the relationship between quality standards and performance measures and the priorities for developing both. The last two proposals in the Priorities and Uses section deal with the role of skill assessment and testing systems as a performance measure and the need for a criteria for the selection and development of tests.

The proposals in the Criteria for Skill Assessment Tests focus on important considerations for the selection or development of a skill testing systems for adult learners. This second group of proposals is based on the background discussion in Chapter 3.

4.1

Priorities and Uses

The first four proposals are concerned with decisions that need to be made before skill assessment and testing systems can be developed or selected (Exhibit 4.1).

Exhibit 4.1
PROPOSALS DEALING WITH
PRIORITIES AND USES

-
- (1) Develop Performance Measures in Response to Quality Standards
 - (2) Prioritize Development of Quality Standards and Performance Measures
 - (3) Use Skill Assessment Tests as the Primary Performance Measure
 - (4) Develop a Common Criteria for Evaluating all Skill Assessment and Testing Systems
-

These four proposals are described in this section.

DEVELOP PERFORMANCE MEASURES
IN RESPONSE TO QUALITY STANDARDS

It is important to clarify the goals of programs before developing performance measures.

It is important to clarify the goals of programs before developing performance measures. The learning goals of educational programs are commonly specified in state quality standards and curriculum frameworks. These standards are an essential prerequisite for development of performance measures.

Proposal 1
DEVELOP PERFORMANCE MEASURES
IN RESPONSE TO PROGRAM QUALITY STANDARDS

Develop Quality Standards to establish goals for adult education and training programs, then use these Quality Standards to guide development of Performance Measures:

- Review and isolate common quality standards for all education providers participating in California Adult Education Strategic Plan.
- Use the common goals as guidelines for developing performance measures for specific program areas.

Clear linkage of Performance Measures to the goals specified in Quality Standards should be given priority.

Quality standards for all education providers participating in California's Adult Education Strategic Plan need to be developed and reviewed in order to isolate a set of common learning goals.¹ These common goals can then be used as guidelines for specifying performance measures in each of the ten publicly funded areas of adult education.

Agency coordination and agreement is important for the effective establishment of these quality standards and performance measures. Achieving common standards and performance measures will pave the way for articulation among programs and open the door to cost-saving collaboration in developing Performance Measures.

PRIORITIZE DEVELOPMENT OF COMMON STANDARDS AND PERFORMANCE MEASURES

Highest priority should go to developing standards and performance measures for basic skills and literacy.

Since there are many different types of adult education classes and programs in California, it is necessary to decide the order in which quality standards and performance measures are developed for each of the ten authorized areas of publicly funded adult education. Highest priority should go to developing standards and performance measures for basic skills and literacy.

Proposal 2

PRIORITIZE DEVELOPMENT OF COMMON STANDARDS AND PERFORMANCE MEASURES

The order of priority for development of common outcome and performance measures should be as follows:

- Basic Skills and Literacy (Elementary and Secondary Basic Skills, English-as-a-Second-Language).
- Other authorized areas of instruction, including Citizenship, Health and Safety, Parenting, Older Adults, Substantially Handicapped, Home Economics, Apprenticeships, or Vocational Education.

Where possible, efforts should be made to develop standards and measures simultaneously for several areas.

¹ Statements of learning goals vary greatly among educational providers. Generally, educational providers that are funded by state apportionments have the most specific and comprehensive statements of learning goals (e.g., Adult Schools, Community Colleges and ROC/Ps). The strategic planning process for quality standards and performance measures will review and integrate learning goals from all providers. However, priority will be given to goals established by providers funded by state apportionment.

To the extent that research and advisory activities can be coordinated, quality standards and performance measures should be developed simultaneously for several instructional areas.

PLACE HIGH PRIORITY ON SKILL ASSESSMENT TESTS AS PERFORMANCE MEASURES

The best common measure of program performance is that programs teach what students need to learn and that students learn what programs teach. Skill assessment tests should be selected or developed to provide a standard common measure of learning achievement among programs and agencies. These tests should be the core, but not necessarily the only measure of performance.

Skill assessment and testing systems should be developed and used to produce a common measure of learning achievement as the primary performance measure of participating agencies.

Proposal 3

PLACE HIGH PRIORITY ON SKILL ASSESSMENT TESTS AS PERFORMANCE MEASURES

Skill assessment and testing systems should be developed and used along with other measures of performance to produce a common measure of learning achievement as the primary performance measure of participating agencies. While the nature of skill testing may vary among areas of authorized instruction, every effort should be made to develop a standard test of learning achievement for each area of instruction. These tests should be the core, but not necessarily the only measure of performance.

Selected skill assessment and testing systems should be capable of measuring both the level of skills and improvement of skills.

DEVELOP GENERAL CRITERIA FOR EVALUATING SKILL ASSESSMENT AND TESTING SYSTEMS

There is a need to use skill assessment and testing systems that meet key needs and produce reliable test data. General criteria should therefore be developed to guide the selection or development of skill testing systems. These criteria should reflect the purposes of testing and professional standards for test design and administration.

General criteria should be developed to guide the selection or development of skill assessment and testing systems.

Proposal 4
DEVELOP GENERAL CRITERIA FOR EVALUATING
SKILL ASSESSMENT AND TESTING SYSTEMS

General criteria should be developed to guide the selection or development of skill assessment and testing systems for applicable authorized areas of public adult education:

- 4-1 Availability of Technical Documentation
- 4-2 Applicability of Testing Systems to Multiple Uses
- 4-3 Pertinence of Scores and Range of Measurements
- 4-4 Documentation of Validity and Reliability
- 4-5 Collection of Representative and Comparative Data
- 4-6 Common Measurement of Skill Levels
- 4-7 Capacity for Reporting and Analysis
- 4-8 Safeguards to Protect Individual Rights and Privacy
- 4-9 Administrative Effectiveness and Efficiency
- 4-10 Ongoing Review and Development

These criteria should address the relevance of testing systems to user needs and professional standards for test design and administration.

The next section of this chapter elaborates each of the proposed general criteria.

4.2

Criteria for Skill Testing Systems

The remaining proposals deal with a general criteria for selecting or developing skill assessment tests for all areas of adult education and training. These proposals provide ten elaborations of Proposal 4 (Exhibit 4.2).

Exhibit 4.2

PROPOSED CRITERIA FOR EVALUATING FOR SKILL
ASSESSMENT AND TESTING SYSTEMS

- 4-1 Availability of Technical Documentation
- 4-2 Applicability of Testing Systems to Multiple Uses
- 4-3 Pertinence of Scores and Range of Measurements
- 4-4 Documentation of Validity and Reliability
- 4-5 Collection of Representative and Comparative Data
- 4-6 Common Measurement of Skill Levels
- 4-7 Capacity for Reporting and Analysis
- 4-8 Safeguards to Protect Individual Rights and Privacy
- 4-9 Administrative Effectiveness and Efficiency
- 4-10 Ongoing Review and Test Development

It is unlikely that any existing test will fully meet all ten of these criteria. However, it is important to establish a comprehensive criteria to guide efforts to develop the best possible testing system.

These ten elaborative subproposals are presented in the following pages.

AVAILABILITY OF TECHNICAL DOCUMENTATION

It is important that skill assessment tests be accompanied by technical documentation that explains the purposes and uses of the test, how the test was developed, and the limitations of the test.

Skill assessment and testing systems should be accompanied by thorough technical documentation.

Proposal 4-1 AVAILABILITY OF TECHNICAL DOCUMENTATION

Skill assessment and testing systems should be accompanied by thorough technical documentation:

- Documentation should cover all major issues of interest to test users (see remaining proposals of this chapter).
- Documentation should be readily available and understandable to users, with summary documents for lay users and detailed documentation for specialists.

The vendors or agencies providing the test should also provide staff to explain or elaborate technical documentation.

Technical documentation should be readily accessible and understandable to users and staff. Detailed documentation should be available to provide elaboration.

APPLICABILITY OF TESTS TO MULTIPLE USES

Where possible, skill assessment and testing systems should be applicable to multiple uses. These uses should include counseling and feedback to learners, placement and certification, and program review and evaluation.

Where possible, skill assessment and testing systems should be applicable to multiple uses. These uses should include counseling and feedback to learners, placement and certification, and program review and evaluation.

Proposal 4-2
SKILL ASSESSMENT TESTS
FOR MULTIPLE PURPOSES

Adopt skill assessment tests that can be used for a variety of purposes within diverse education and training settings. The tests should be applicable to the following uses:

- Counseling and Feedback to Users
- Placement and Certification
- Program Review, Monitoring and Evaluation

It is particularly important that the same test, or crosswalk-linked tests, be used to evaluate learner skills and program performance.

If it is not possible or desirable to use the same test for multiple purposes, test scores should be translatable to core scores that provide a common frame of reference for multiple applications.

PERTINENCE OF SCORES
AND RANGE OF MEASUREMENTS

Skill assessment and testing systems should have scoring scales and ranges of measurement pertinent to multiple user needs. The types of skills measured should be linked to quality standards specifying the types of skills to be taught in programs. The level and range of skills measured by the test should be appropriate to the programs using the test. The questions and scoring systems should be unbiased to subgroups of learners. Finally, the scoring system of the tests should provide a meaningful frame of reference for learners and programs.

The scoring system and range of measurement should provide a commonly understandable frame of reference for a variety of groups using test scores.

Proposal 4-3**APPROPRIATE MEASUREMENT SCORES**

Skill assessment and testing systems should have measurement and scoring systems that meet the needs of test users:

- Tests should measure skills that are derived from program quality standards specifying the types of skills to be taught in programs.
- The level and range of skill measured by the test should match the level and range of skills taught by programs using the test.
- Test questions and measurements should not be biased in favor or against subgroups taking the test.
- The scoring system and range of measurement should provide a commonly understandable frame of reference for a variety of groups using test scores.

If more than one test is used for a given skill area, these tests should be linked by crosswalks to allow development of common scores.

Test users should explore all possible approaches for ensuring that tests have appropriate measurement scores. Possible approaches include crosswalking scores to translate test results into common scales, use of general and detailed tests, and options for reporting test results in fashions appropriate to specific uses (e.g. criterion-based and norm-based scores).

**DOCUMENTATION OF VALIDITY
AND RELIABILITY**

Skill assessment and testing systems should accurately measure the skills that they claim to measure, and they must measure skill levels reliably over time. A test reported to have validity in measuring general math skill should not measure only multiplication skills. Equally important, skill levels measured at one point in time should not be measured at a different level at another point in time if the skills of the test taker have not changed.

Expected ranges of error in validity and reliability should be specified in accord with professional standards for testing and measurement.

Proposal 4-4
DOCUMENTATION OF VALIDITY AND RELIABILITY

Skill assessment and testing systems should be accompanied by rigorous documentation of validity and reliability:

- Testing system should accurately measure what they are intended to measure (e.g. a math test should measure math skill within a specified range of competency, not computer skills).
- Testing systems should measure specified skills consistently over time (e.g. a test taker who scored at one level one week should not have significantly lower scores when taking the same test two weeks later).

Documentation of validity and reliability should clearly specify the ranges of error expected under different testing conditions. The reported ranges of error should be acceptable to test users.

Expected ranges of error in validity and reliability should be specified in accord with professional standards for testing and measurement. Test users should select or develop testing systems that have acceptable ranges of error for intended uses.

**COLLECTION OF
REPRESENTATIVE AND
COMPARATIVE DATA**

Data from skill assessment and testing systems should be collected in ways that ensure that it is representative of learner populations in programs using skill tests, and that test data can be compared to reference groups and among subgroups of the tested population.

Test score data should be collected and stored in ways that ensure that it is representative of target populations and capable of producing comparative analysis.

**Proposal 4-5
COLLECTION OF REPRESENTATIVE
AND COMPARATIVE DATA**

Test score data should be collected and stored in ways that ensure that it is representative of target populations and capable of producing comparative analysis with reference groups and among subgroups of test takers:

- Test data should be collected from all learners or a random sample of learners to ensure that aggregate scores are representative of programs and specified groups of learners.
- Test data should be collected from comparison groups (e.g. total adult population) and with demographic data on test takers (e.g. age, race, sex, socioeconomic status) to allow comparisons among test takers and development of test score norms.
- Pre/post tests should be administered to program participants to ensure that test scores measure skill improvement over time.

The above specifications should be implemented in ways that meet professional standards of sampling and measurement.

Collection of test data that does not meet acceptable professional sampling and measurement standards may do more harm than good.

**COMMON MEASUREMENT
OF SKILL LEVELS**

It is important that assessment tests provide measurement of skills that can be compared to other populations and the scores of other tests. Where possible, crosswalks should be developed or used to translate test scores to scales that are comparable with other tests.

**Proposal 4-6
LINKAGE TO OTHER TESTS
AND MEASUREMENTS**

Skill assessment tests should be selected or developed for their capacity to use crosswalks that allow scores to be translated and compared to the scores of other tests. If several tests are used to measure a given area of skills, crosswalks should be developed to allow a common interpretation of scores from different tests. Ranges of error caused by translation of scores should be clearly specified.

When crosswalks are used to translate test scores, care should be taken to assess and recognize the ranges of error to be expected from the translations.

CAPACITY FOR REPORTING AND ANALYSIS

The standard reports produced by a testing system should be pertinent to all the multiple uses intended for the test.

The standard reports produced by a testing system should be pertinent to all the multiple uses intended for the test, and test data should be collected and stored to provide options for special use and analysis.

Proposal 4-7 CAPACITY FOR REPORTING AND ANALYSIS

The standard reports provided by skill assessment tests and options for use and analysis of test data should meet a wide range of existing and potential user needs:

- Standard reports produced by the testing system should meet all the routine needs of users.
- Test data should be collected and stored to maximize existing and prospective needs for analysis by storing data on individuals rather than group batches, collection of demographic and program affiliation data on test takers, and linkage of individual pre/post test scores.
- The *EduCard_{TM}* and other procedures should be used to make test data available to learners and other authorized persons.

These and other procedures should be implemented in ways that protect the privacy of individuals and programs.

Test takers should have easy access to their scores; and authorized program staff should have test data readily available for purposes of counseling, certification, and analysis for program evaluation.

SAFEGUARDS TO PROTECT INDIVIDUAL RIGHTS AND PRIVACY

Storage and access of test data should be organized in ways that protect individual rights and privacy. Individuals should have the right to review their own test scores with minimal effort. Additionally, access to test scores must be restricted to authorized persons.

Skill testing systems should organize test data and storage provisions to provide maximum privacy and rights to individuals.

**Proposal 4-8
SAFEGUARDS TO PROTECT
INDIVIDUAL RIGHTS AND PRIVACY**

Skill testing systems should organize test data and storage provisions to provide maximum privacy and rights to individuals. Key provisions should include:

- Test takers should be given a full and understandable explanation of the purposes for testing.
- Test takers should be given easy access to review their own test scores and to make appeals for retesting.
- Test score data should be stored in ways that make files and data accessible only to authorized persons.
- Test score reports should be prepared in ways that do not disclose inappropriate information about individual test takers.

These and other procedures should be developed and reviewed regularly in public meetings allowing input from all concerned parties.

Procedures to protect individual rights and privacy should be developed and periodically reviewed in open forums allowing input from all interested persons.

**COST EFFECTIVENESS AND
ADMINISTRATIVE EFFICIENCY**

Skill assessment and testing systems should be selected and developed for their cost effectiveness and administrative efficiency. Administrative efficiency requires that testing systems be selected that have the lowest total cost among those systems that meet the previously discussed criteria. Total cost includes: (1) acquisition and development cost, (2) provider training cost, (3) proctoring cost, (4) evaluation cost, (5) data storage cost, (6) cost of scoring and interpretation, and (7) the cost of consultant specialists. Attention should also be given to the selection and development of tests that can be administered in accord with acceptable standards by test users.

Skill testing systems should be selected or developed for their cost effectiveness and administrative viability.

Proposal 4-9
COST EFFECTIVENESS AND
ADMINISTRATIVE EFFICIENCY

Skill testing systems should be selected or developed for their cost effectiveness and administrative viability. Skill assessment tests should have the lowest possible cost within the requirements of the selection and development criteria (e.g. the costs of using the test should be lower than the human and organizational cost of not using the test). Costs to be considered should include acquisition and development, provider training, proctoring, evaluation, scoring and analysis, consultants, and data storage.

All in all, the costs of skill assessment tests should be lower than the human and organizational costs of not using the tests.

ONGOING REVIEW
AND DEVELOPMENT

Skill assessment tests should be selected for their long-term capacity for ongoing development and improvement.

Proposal 4-10
ONGOING REVIEW AND DEVELOPMENT
Test providers should provide evidence of ongoing commitment to review, evaluation, and test improvement:

- Test providers should openly provide test data for review and assessment by users.
- Tests should have regular third party evaluations by qualified parties selected for their independent objectivity, and capacity to provide balanced and comprehensive evaluations.
- Tests providers should demonstrate a history and ongoing commitment to research and development.

These and other commitments to review and improve tests should indicate a spirit of ongoing, cooperation and responsiveness to test users.

Test providers should give evidence of regular third party evaluations and ongoing research and development.

Test providers should give evidence of regular third party evaluations and ongoing research and development.

They should also share information about their test and respond to user needs.

The ten proposals above compose preliminary criteria for selecting or developing skill assessment and testing systems. These recommended criteria are distillations of elaborations provided in Chapter 3. Many criteria may need further elaboration, and additional criteria may be needed.

Chapter 5

Future Action

This chapter provides a general outline of research and planning actions necessary for the implementation of the proposals presented in this paper.

This chapter provides a brief summary of the issues reviewed in the four prior chapters and a summary of the future actions in research and planning needed to implement the proposals made in this report.

5.1

General Review

Program quality standards are a necessary prerequisite for the establishment of performance measures. It is important that these standards define the minimum requirements, the normal expectations and the conditions of excellence to ensure that the adult education system meets the needs of stakeholder. These standards should document measures of performance that focus on the skills that students will need for work and life, and the skill achievement expected to result from participation in instructional programs.

Skill assessment tests are recommended as the primary mechanism for measuring student outcomes.

General criteria are presented for selecting or developing skill assessment and testing systems. It is proposed that first priority be given to developing program quality standards and measures that are designed to measure basic skills and literacy for adult learners. The proposals in the report are designed to facilitate the selection or development of skill assessment and testing systems that can be used in both traditional classroom and non-classroom settings.

Ultimately, the proposed criteria should be used for selection or development of skill assessment and testing systems for all adult education programs. Where possible, scores from these testing systems should be used to guide students, certify skill attainment, and to evaluate the quality of the programs.

5.2

Developmental Considerations

When the proposals of this report have been reviewed, modified and approved via advisory committees and agency heads, attention must be directed to the issue of how to put them into action.

It is important to recognize the need to have consistency in the implementation of these proposals. It is also important to recognize that many of the proposals put forward in this report are interdependent. Failure to apply one proposal may undermine the successful application of others. Because high quality testing and assessment requires careful testing, management and analysis, there is a vital need for consistent and systematic work to establish the testing system that will provide the necessary results for all stakeholders.

It is vital that the testing system be seen as legitimate at the local level. This suggests that the testing system must be developed and maintained through an ongoing process of state-local collaboration.

When looking to the future, it is necessary to suggest the types of actions that may flow from the proposals presented in this report. The first alternative might be to select the tests as they currently are used. The second alternative is to alter existing testing systems to ensure that they are applicable to the needs of the adult education system. A third and admittedly unlikely alternative, would be to develop new testing systems.

In all of these categories there is a need to ensure that there is close integration of quality standards and performance measures. Thus, there is a need to ensure that the two procedures for development have a close working arrangement in the future.

Another area that requires work and creative innovation is the use or development of non-test performance measures. Such non-test performance measures may include student retention rates, surveys of student satisfaction, and certifications from professionally acceptable educational institutions.

Ultimately, it is vital that the testing system be seen as legitimate at the local level. This suggests that the testing system must be developed and maintained through an ongoing process of state-local collaboration.

There should be an increase in the level of test assessment and design. This is especially true for the development of multiple use tests and for the use of tests in nontraditional areas. Finally, there is a real need for a high quality tracking design of those who leave the educational system early. Work must also be continued to develop more effective and representative tests, and more efficient management and proctoring in the field.

5.3

Future Actions

This report is the first step in a multi-stage initiative to develop consistent Quality Standards and Performance Measures for adult education.

Future planning and implementation will be conducted under the guidance of the Subcommittee on Quality Standards and Performance Measures of the Interim Adult Education Steering Committee. This work will be subject to the review and approval of State Superintendent of Public Instruction Bill Honig, State Community College Chancellor David Mertes, and other participating agency heads.

Approved priorities and criteria from this paper will be used in conjunction with agency developed program quality standards to review and recommend one or more skill assessment and testing systems for appropriate authorized areas of adult education. It is proposed that first priority be given to developing quality standards, skill testing systems, and other non-test performance measures for areas of instruction dealing with basic skills and literacy.

Bibliography

K.A. Brownlee, *Statistical theory and methodology in science and engineering*. Wiley, New York, 1960.

E. Brunswik, *Perception and the representative design of psychological experiments*. (2nd ed.) Berkeley: University of California Press, 1956.

Donald T. Campbell and Julien C. Stanley, *Experimental and Quasi-Experimental Designs for Research*. Rand McNally College Publishing Company, Chicago, 1963.

Donald T. Campbell, *Psychology Bulletin*, "Factors relevant to the validity of experiments in social settings". 1957, 54, 297-312.

Donald T. Campbell, *American Psychologist*, "Recommendations for APA test standards regarding construct, trail, or discriminant validity". 1960, 15, 546-553.

V.R. Cane and A.W. Heim, *Quart. Journal. exp. Psychology*. "The effects of repeated testing: III. Further experiments and general conclusions". 1950, 2, 182-195.

G.A. Ferguson, *Statistical analysis in psychology and education* New York: McGraw Hill, 1959.

R.A. Fisher, *Statistical methods for research workers* (1st ed.) London: Oliver and Boyd, 1925.

R.A. Fisher, *The design of experiments* (1st ed.) London: Oliver and Boyd, 1935.

C.Y. Glock, *The language of social research*. "Some applications of the panel method to the study of social change. In P.F. Lazarsfeld and M. Rosenberg (Eds.)". Glencoe, Ill.: Free Press, 1955. Pp. 242-249.

C.Y. Glock, *The study of short run social change, in preparation*. "The effects of re-interviewing in panel research 1958. Multilith of a chapter to appear in P.F. Lazarsfeld (Ed.)".

O. Kempthorne, *The design and analysis of experiments*. New York: Wiley, 1952.

O. Kempthorne, *Journal American. Statist. Association*. "The Randomization theory of statistical inference". 1955, 50, 946-967; 1956, 51, 651.

R.E. Lana, *Psychology Bulletin*. "Pretest-treatment interaction effects in attitudinal studies". 1959, 56, 293-300. (a)

E.F. Lindquist, *Statistical analysis in educational research*. Boston: Houghton Mifflin, 1940.

W.A. McCall, *How to experiment in education*. New York: Macmillan, 1923.

E.B. Page, *Journal education. Psychology*. "Teacher comments and student performance: A seventy-four classroom experiment in school motivation". 1958, 49, 173-181.

J.C. Reed, *Journal Education Psychology*. "Some effects of short term training in reading under conditions of controlled motivation". 1956, 47, 257-264.

H. Scheffe, *Ann. math, Statist*. "Alternative models for the analysis of variance". 1956, 27, 251-271.

H.L. Smith and H. Hyman, *Published. opinion. Quart*. "The biasing effect of interviewer expectations on survey results". 1950, 14, 491-506.

M.G. Sobol, *Journal American Statist. Association*. "Panel mortality and panel bias". 1950, 54, 52-68.

R.L. Solomon, *Psychology Bulletin*. "An extension of control group design". 1949, 46, 137-150.

Standards for Educational and Psychological Testing, American Psychological Association, Washington, D.C., May 1986.

J.C. Stanley and Ellen Y. Beeman, *Psychology Newsletter*. "Restricted generalization, bias, and loss of power that may result from matching groups". 1958, 9, 88-102.

B.J. Winter, *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Recruitment admissions and handicapped students. A guide for compliance with Section 504 of the Rehabilitation Act of 1973.

"American Association of Collegiate Registrars and Admissions Officers and American Council on Education. (1978, April). Admissions tests". Washington, DC.

Responsibilities of users of standardized tests. "American Personnel and Guidance Association. (1980)". Falls Church, VA.

Technical recommendations for psychological tests and diagnostic techniques. "American Psychological Association. (1954)". Washington, DC.

American Psychologist, 21 (12), 1141. "American Psychological Association. (1966a). Automated test scoring and interpretation practices. In Proceedings of the American Psychological Association".

Standards for educational and psychological tests and manuals. "American Psychological Association. (1966b)". Washington, DC.

Standards for educational and psychological tests. "American Psychological Association. (1974)". Washington, DC.

American Psychologist. "American Psychological Association. (1977). Guidelines for nonsexist language in APA journals", 32 (6), 487-494.

American Psychologist, 32. "American Psychological Association. (1977). Standards for providers of psychological services". (6), 495-505.

American Psychologist 36 "American Psychological Association. (1981). Ethical principles of psychologists". (6), 633-638.

Ethical principles in the conduct of research with human participants. "American Psychological Association (in press)". Washington, DC.

Statement on legislation affecting testing for selection in educational and occupational programs.

"Association for Measurement and Evaluation in Guidance". (1980, October).

Guidelines on the uses of College Board test scores and related data. "College Entrance Examination Board." (1981). New York.

Guidelines for use of computerized testing services. "Colorado Psychological Association". (1982, January).

Federal Register, 42. :Department of Health, Education, and Welfare. (1977). Nondiscrimination on basis of handicap in programs and activities receiving or benefiting from federal financial assistance". (86), 22676-22702.

Principles for the validation and use of personnel selection procedures (2nd ed.). "Division of Industrial and Organizational Psychology" (1980). Berkeley, CA: American Psychological Association.

Education for all Handicapped Children Act of 1975, 20 U.S.C. § 1401-1461 (1976).

Standards for quality and fairness. "Educational Testing Services" (1983). Princeton, NJ.

Federal Register, 43. "Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. (1978). Adoption by four agencies of uniform guidelines on employee selection procedures". (166), 38290-38315.

Federal Register, 44. "Equal employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of Labor, and Department of the Treasury. (1979). Adoption of questions and answers of clarify and provide a common interpretation of the uniform guidelines on employee selection procedures". (43), 11996-12009.

Federal Register, 45. "Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of the Treasury, and Department of Labor Office of Federal Contract Compliance Programs. (1980). Adoption of additional questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures". (87), 29530-29531.

New directions for program evaluation.

"Evaluation Research Society, Inc. (1982).
Standards for program evaluation". In P.H. Rossi
(Ed.), (pp. 37-48). San Francisco: Jossey-Bass.

K. Heller, W. Holtzman and S. Messick, *Placing
children in special education: A strategy for
equity*. (1982) Washington, DC: National
Academy Press.

*Report of the joint AERA, APA, NCME,
committee for review of the standards of
educational and psychological tests.* "Joint
AERA, APA, NCME Committee for Review of
Standards for Educational and Psychological
Tests". (1979). Washington, DC.

*Standards for evaluations of educational
programs, projects, and materials.* "Joint
Committee on Standards for Educational
Evaluation" (1981). New York: McGraw-Hill.

*NCME statement on educational admissions
testing.* "National Council on Measurement in
Education Board of Directors". (1980), April).
Washington, DC.